Hybridisation and phylogenomics of Betula L. (Betulaceae).

Wang, Nian

Hybridisation and phylogenomics of *Betula* L. (Betulaceae)

Nian Wang

School of Biological and Chemical Sciences,

Queen Mary University of London,

Mile End Road,

London E1 4NS

Supervisor: Dr Richard J. A. Buggs

A thesis submitted to the

University of London

for the degree of

Doctor of Philosophy

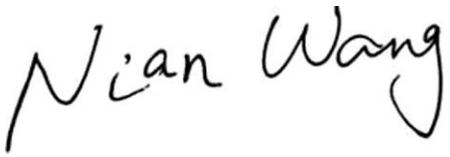
November 2015

Declaration

I certify that this thesis and the research conducted here are the product of my own work, and that all ideas and quotation from other works, published or otherwise, are fully acknowledged in accordance with standard refereeing practices in the biological sciences.

I acknowledge particular data acquisition and analytical assistance as follows:

James Borrell helped do Ecological niche modelling based on which I produced Figure 2.4. Richard Nichols helped develop scripts to do cline analysis using the mixed-effect model, based on which I produced Figure 2.3 in R. Jasmin Zohren helped to write an R script to change mapped loci name for Chaper 5.



Nian Wang

Acknowledgements

First and foremost I would like to express my sincere gratitude to my PhD supervisor Richard J. A. Buggs. Richard, as an outstanding researcher and supervisor, has made my research as smooth as possible. Thanks to him for his guidance, inspiration, patience and weekly discussion, which have benefitted me not only for my PhD at QMUL but also for my career. I owe a big thanks to Prof. Andrew R. Leitch, whose ideas in science are amazing.

I extend my sincere thanks to Hugh McAllister, whose knowledge on the taxonomy of birch is second to none, his comments and guidance have been insightful and invaluable. Many thanks to Paul Bartlett, who manages a great Botanic Garden with an excellent collection of birches, for his generosity in providing me with almost all samples I need. I wish to thank Prof. Richard Nichols, as one of my panel members and the organiser of the 'Labchat', whose comments on my project are insightful. I wish to acknowledge my sincere gratitude to James Borrell for helpful discussions on sciences and help in data analysis; to Jasmin Zohren and Maité Guignard for their help on R scripts; to Laura Kelly for her insightful comments on Chapter 4; to Monika Struebig, who has provided excellent support during the period of my study and has made life in the lab much easier.

I also would like to thank the following people in alphabetic order: Hannes Bechler, Peter Brownless, Endymion Cooper, William Crowther, Paul Grogan, Hao-Chih Kuo, Ilia Leitch, Lu Ma, Xiu-Guang Mao, Jaume Pellicer, Martyn Rix, Elizabeth Sollars, Jonathan Stocks, Bruno Vieira, Gemma Worswick, Zhi-Kun Wu and Jie Zeng for various helps.

Last but not least I would like to thank my family members: Pei-Gang Wang, Shu-Ping Zhai and Xin-Ying Zhao for their continued support and encouragement.

Abstract

Hybridisation and polyploidy are important in the evolution of species. The genus *Betula* L. is an ideal model to study these processes as species of this genus hybridise frequently and polyploid species are common. In this thesis, I investigated the hybridisation of three *Betula* species in Britain; the phylogenetic relationships and genome size of all known *Betula* species, and conducted phylogenomic analysis of diploid *Betula* species.

A cline of introgression of microsatellite marker alleles from *B. nana* was found extending into *B. pubescens* populations far to the south of the current *B. nana* range in Britain. We suggest that this genetic pattern is a footprint of a historical decline and/or northwards shift in the range of *B. nana* populations due to climate warming in the Holocene.

The Atkinson Discriminant Function (ADF) is a leaf-morphology metric to distinguish between *B. pubescens* and *B. pendula*. We test it on 944 trees sampled across Britain against species' discriminations made using 12 microsatellite loci and found that the accuracy of the ADF can be raised to 97.5% by using an ADF of -2 rather than zero as the boundary line between the species.

The taxonomy of the genus *Betula* is debated and a new classification has been proposed in a recent monograph. We evaluated it using ITS and restriction site associated DNA sequencing (RADSeq). The result based on ITS largely agrees with species classification in the recent monograph but with uncertainties. Phylogenomic analysis of 587 RAD loci for *Betula* diploid species using coalescence-based methods, a concatenation method and binary presence/absence of RAD loci produced well-resolved trees with similar topology. Based on these analyses, we propose a new classification of *Betula* into four subgenera and seven sections. Further research is needed to infer the parental origins of polyploid species within *Betula*.

Contents

List of Figures

List of Tables

Chapter 1 General Introduction

Introduction to relevant theory

The importance and outcomes of hybridisation

Hybridisation is commonplace in plants and animals with about 25% and 10% of plants and animal species being estimated to hybridise with their relatives (Mallet, 2005). Although the role of hybridisation in evolution remains controversial (Mayr, 1963; Arnold *et al.*, 1999), its importance has been increasingly accepted (Anderson, 1949; Anderson, 1953; Stebbins, 1959; Lewontin & Birch, 1966; Dowling & Demarais, 1993; Dowling & Secor, 1997). Hybridisation may lead to speciation in two main ways: homoploid hybrid speciation (Buerkle *et al.*, 2000; Mallet, 2007) and allo-polyploidy (Otto & Whitton, 2000; Ramsey & Schemske, 2002; Mallet, 2007). In addition to the positive roles hybridisation plays in evolution, it has some negative effects. For example, hybridisation can cause species extinction via genetic assimilation (Levin *et al.*, 1996; Rhymer & Simberloff, 1996; Huxel, 1999), pollen swamping (Buggs & Pannell, 2006; Prentis *et al.*, 2007) or replacement of parental species by a more vigorous hybrid (Wolf *et al.*, 2001). Hence, it can pose a threat to endangered species. Many native species may be under threat by hybridisation especially in the context of global warming, which can bring previously geographically separated species into contact.

Introgression is the transfer of genetic materials from one species into another via repeated backcrossing of hybrids. It occurs frequently in plants and animals and plays an important role in evolution by introducing genetic variation (Abbott *et al.*, 2013), which can be adaptive due to the transfer of advantageous alleles and can provide the basis for natural selection to act on (Rieseberg, 2001; Seehausen, 2004; Genner & Turner, 2012). Adaptive introgression has been documented in numerous plants and animals, such as in *Iris* (Martin *et al.*, 2006), *Helianthus* (Vekemans, 2010) and *Heliconius* (Pardo-Diaz *et al.*, 2012) and it may facilitate adaptation and expansion to novel habitats (Whitney *et al.*, 2010).

Introgression can be bidirectional (Stewart *et al.*, 2012; Clark *et al.*, 2015) or unidirectional (Trucco *et al.*, 2009; Thompson *et al.*, 2010; Balao *et al.*, 2015).

Unidirectional introgression is a potential threat to rare species. Several mechanisms have been suggested to be responsible for unidirectional introgression, such as the relative abundance of hybridising species (Burgess *et al.*, 2005; Lepais *et al.*, 2009), natural selection (Plotner *et al.*, 2008; Fitzpatrick *et al.*, 2010), spatial expansion of one species (Currat, 2008), cytonuclear incompatibility (Bolnick *et al.*, 2008), ploidy level difference (Stebbins, 1971) and unilateral interspecific incompatibility (Lewis & Crowe, 1958). Empirical and simulation studies have shown that unidirectional introgression often occurs from local species to the invading species leaving a cline of introgression behind (Buggs, 2007; Currat *et al.*, 2008). Such a cline of introgression can be used to trace the past demography of the hybridsing species.

Incomplete lineage sorting (ILS)

Ancestral populations harbour allelic diversity, which originally arose from mutations. However, only a subset of alleles passes on from these ancestral populations to the present lineages by chance (genetic drift) or natural selection. This may result in a coalescent history where alleles do not coalesce (looking backwards in time) into an ancestral allele until times deeper than speciation events. This is called incomplete lineage sorting (ILS) (Maddison, 1997). For example, in Fig. 1.1, in the ancestral populations of species A, B and C, there are three alleles marked in green, blue and yellow, respectively. Allele B in the lineage leading to species B failed to coalesce in the recent T1 speciation event but coalesced with allele C before the previous T2 speciation event. This ILS may result in gene tree and species tree discordance (see below).

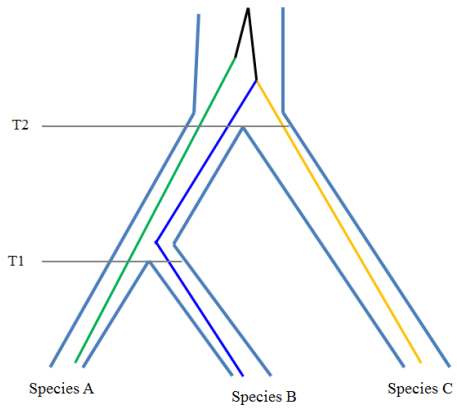


Figure 1.1 A schematic representation of incomplete lineage sorting (ILS). T1 and T2 indicate speciation events.

Allele sharing

Allele sharing is often detected between species, which can arise from ILS (Muir & Schlotterer, 2005; Tsuda *et al.*, 2015), introgression (Salzburger *et al.*, 2002) or horizontal gene transfers (HGTs) (Syvanen, 2012). ILS most commonly occurs when an ancestral species undergoes several speciation events in a short period of time. It is more likely to occur for species with a large effective population size (Degnan & Rosenberg, 2009). Introgression mostly occurs in hybridising species whereas HGT commonly occurs in bacteria (Wiedenbeck & Cohan, 2011; Polz *et al.*, 2013) and occasionally in distantly related species (Won & Renner, 2003; Fuentes *et al.*, 2014; Wang *et al.*, 2015). It is very difficult to distinguish between ILS and introgression. Several methods have been proposed to distinguish the two. For example, ILS has similar geographical patterns of genetic admixture across the range of a species and can happen between species with reproductive isolation. Introgression is only possible for hybridising species and higher levels of genetic admixture based on a single locus or multiple linked loci would be expected for sympatric populations than allopatric populations (Twyford & Ennos, 2012). Hence, a comprehensive sampling of species is preferable to distinguish ILS from introgression based on geographical signalling.

Hybrid zones

The term "hybrid zone" refers to a geographic region where genetically distinct species or populations meet and produce hybrids (Barton & Hewitt, 1985). This can cause a "cline" where there is a gradient of phenotypic or genotypic frequencies from one species to another species. Hybrid zones have been documented in numerous plants and animals (Petit *et al.*, 1999; Rieseberg *et al.*, 1999; Nielsen *et al.*, 2004; Buggs, 2007). Hybrid zones can be formed due to the initial genetic divergence of adjacent populations, which may eventually lead to parapatric speciation or be formed by secondary contact of divergent lineages of allopatric distributions.

If hybrids have lower fitness compared with their parental genotypes, natural selection may eliminate them. In this situation, hybrid zones are maintained by natural selection against hybrids and the on-going dispersal of gametes. This type of hybrid zone is called a tension zone (Slatkin, 1973; Barton & Hewitt, 1985). In another case, hybrids are more vigorous than their parental types in certain environments, such as Moore's bounded hybrid superiority zones (Moore, 1977) and mosaic hybrid zones (Harrison & Rand, 1989). Hybrid zones provide a natural window to study the spread of genes

across diverging taxa and to quantify genetic differences underling speciation (Hewitt, 1988; Barton & Hewitt, 1989; Harrison, 1990).

Hybrid zone movements have been extensively studied (Britch *et al.*, 2001; Campbell, 2004; Cruzan, 2005; Buggs, 2007). In some cases, the number of one species decreased as a result of hybridisation with another species, resulting in the shift of hybrid zones. Hybrid zone movement can be rapid especially in the context of climate change.

Species range dynamics

The current distribution of species, though primarily determined by their ecological niches, has been shaped by a combination of factors such as climate change, human introductions (Vitousek *et al.*, 1997; Hewitt, 1999), geographical/geological changes due to tectonic activities and new adaptations to particular soil types and moisture regimes. Most species ranges are therefore dynamic and undergo series of expansions and contractions. It is generally recognised that species retreat into refugia during glaciation whereas they colonise deglaciated areas during inter-glacial periods. Species may have experienced several episodes of range retreat and expansion.

Past range change of species can be traced from herbarium records (Tingley & Beissinger, 2009), ecological niche modelling (ENM) (Peterson, 2003), genetic evidence (Hewitt, 2000) and the fossil record. Herbarium records coupled with field investigation are useful to trace the range shift of a species. In addition, dated fossil pollen and plant macro-fossils have been of crucial importance in providing information on the past range change of species. Recently, ENM has been commonly adopted to predict the past, present and future distribution of a species. ENM, also known as species distribution modelling, uses computer algorithms to predict the distribution of species based on a mathematical representation of their realized ecological niche (Elith & Leathwick, 2009) based on climate data of the localities of a species. ENM has been frequently used in several research areas such as conservation biology (Thorn *et al.*, 2009; Marini *et al.*, 2010), evolution (Jakob *et al.*, 2007) and population demography (Tsuda *et al.*, 2015).

In addition, species range dynamics can be indirectly inferred by comparing the distribution of genetic diversity. Generally, refugial populations tend to harbour high levels of genetic diversity and private alleles, which decrease towards newly colonised habitats due to genetic drift. Species range shifts can also be inferred by studying

hybrid zones. When one species invades the range of a species it hybridises with, a cline of introgression may occur from the local species to the invading species, with a high level of introgression evident near the front of the hybrid zone and less towards its tail (Buggs, 2007). By detecting such a cline, the dynamics of a hybrid zone can be inferred. It is preferable to use multiple lines of evidence to gain a clear picture of past species range shifts.

Pollen records

Reconstructing past environments, especially Holocene environments, has been extensively studied based on various sources such as written records, tree-ring analysis, pollen records and macrofossils. Among these, pollen records are frequently used to infer the type of vegetation and its corresponding climate. Pollen grains and spores can be well preserved in lake muds, peat bogs and other substrates that give anaerobic conditions. Pollen abundance and the plants that produced the pollen grains can be determined. In addition, the age of multiple layers can be dated by stable isotype analysis. Hence, the vegetation type and its abundance in the past can be determined by associating the date of layer with its pollen types and abundance. The European Pollen Database (EPD, http://www.europeanpollendatabase.net/data/) records many pollen sites in European countries with detailed information on pollen type, dates and coordinates of pollen sites.

Species delimitation

The definition of the concept of a species is controversial. Several species concepts have been proposed, such as the phylogenetic species concept (Eldredge & Cracraft, 1980), the evolutionary species concept (Simpson, 1951; Simpson, 1961; Wiley, 1978), the cohesion species concept (Templeton, 1989), the genic species concept (Wu, 2001) and the biological species concept (Mayr, 1942; Mayr, 1963). Each concept has its own limitations (Hausdorf, 2011). Currently, the most popular one used in evolutionary biology is the biological species concept, namely, taxa with complete reproductive isolation or groups of interbreeding natural populations that are reproductively isolated from other such groups.

Species delimitation can be challenging based on morphology as morphological characters are often plastic. For species, like *Betula pubescens* and *B. pendula,* there is no clear boundary but a continuum of leaf morphology. DNA barcoding, a relatively new method to delimitate species (Hebert *et al.*, 2004b; Li *et al.*, 2011), has proved to

be effective in distinguishing species and discovering cryptic species (Hebert *et al.*, 2004a; Kress *et al.*, 2005; Hajibabaei *et al.*, 2006). However, in plants DNA barcoding is still a challenge due to the lack of universal barcodes, the common occurrence of polyploidy, hybridisation and introgression, ancestral polymorphism, presence of paralogs and horizontal gene transfers (Chase *et al.*, 2005; Cowan *et al.*, 2006).

Polyploidy

Polyploidy refers to whole genome duplication (WGS), consisting of two main types: autopolyploidy and allopolyploidy. Autopolyploids have multiple sets of chromosomes derived from a single species whereas allopolyploids have undergone chromosome duplication following inter-specific hybridisation (Soltis & Soltis, 1999; Ramsey & Schemske, 2002). Polyploidy is common in plants, especially in angiosperms. Many important crops are polyploids, such as wheat, maize, cotton and potato (Leitch & Leitch, 2008). It plays an important role in plant speciation and diversification with nearly all extant plant species having experienced multiple rounds of polyploidy in their early history (Jiao *et al.*, 2011).

After polyploidy, the genome seems to be more dynamic. Genes can lose function, gain new function or be expressed in different tissues. For example, multiple gene loss and gain occuring within a few generations after polyploidy has been detected in newly synthetic *Brassica* (Song *et al.*, 1995). Also, in natural systems, such as in *Tragopogon* which were formed in the past 80 years, gene loss and silencing occurs in tissues specifically and non-randomly (Buggs *et al.*, 2009; Buggs *et al.*, 2010; Buggs *et al.*, 2012b; Chester *et al.*, 2012). Established polyploids may have undergone further genomic rearrangement over the long term, resulting in diploidisation (Fawcett *et al.*, 2009).

Phylogeny and Phylogenetics

Phylogeny, a term first coined by Ernst Haeckel, describes the evolutionary history of a group of species. Phylogenetics is the study of the evolutionary relationships of organisms based on morphological data, DNA sequencing data or molecular markers. Phylogenetic studies have been used to address a series of questions in evolutionary biology, such as hybridisation, introgression and phylogeography. Phylogenetics is also frequently used to assess classification of species based on morphological characters (Doolittle, 1999). Generally, multiple individuals of one species are expected to form a monophyletic group which means they are derived from a common ancestor and share derived features (de Queiroz, 2007). Species from a section are

expected to form a monophyletic group which is in turn monophyletic with other sections within the same subgenus. However, monophyly is often violated if hybridisation occurs between closely-related species, which could render paraphyletic relationships (Hörandl, 2006; Hörandl & Stuessy, 2010) (Fig. 1.2). Also, if allopolyploidy occurs, a network relationship between polyploids and their parental species would be expected (Linder & Rieseberg, 2004).

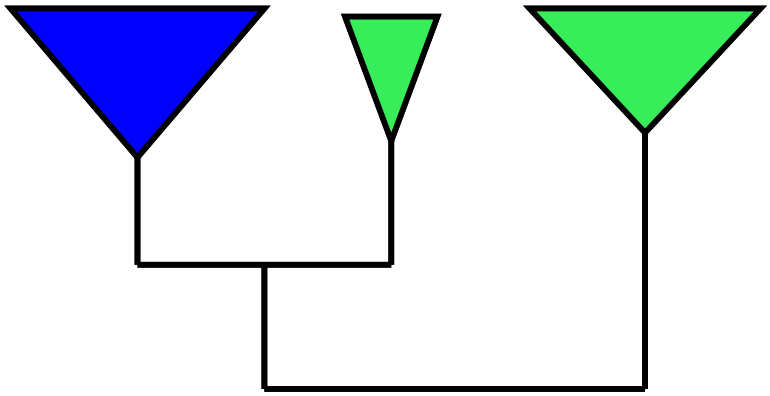


Figure 1.2 A schematic representation of monophyly and paraphyly. The blue and green are monphyletic and paraphyletic, respectively.

Gene trees and species trees

A phylogenetic tree constructed using a single locus is termed a gene tree, which is often discordant with the relationships between species (species tree), due to various factors such as ILS (see above), hybridisation, polyploidy and presence of paralogs. In order to mitigate the effect of ILS, concatenation of multiple loci into a supermatrix has been extensively used to build phylogenetic trees assuming each gene tree has a similar phylogeny (Huelsenbeck *et al.*, 1996; de Queiroz & Gatesy, 2007). However, this method sometimes results in a strongly supported tree which is discordant with the species tree especially when the level of ILS is high (Degnan & Rosenberg, 2009). Recently, coalescent methods for estimating phylogenetic tree have been developed. These methods are based on the idea that each gene tree is a random tree that is generated from the underling species tree. Hence, species trees can be studied in the framework of probabilistic models that takes into account the probability distribution of gene trees (Liu *et al.*, 2008; Liu *et al.*, 2009b).

Phylogeography

Phylogeography is a discipline that studies the geographical distributions of gene genealogies or phylogenetics within species (Avise, 2000). It is commonly used to identify refugial populations (Tremblay & Schoen, 1999; Gavin *et al.*, 2014), infer the origin of species (Olsen & Schaal, 1996) uncover cryptic species (Rissler & Apodaca, 2007) and trace colonisation routes. In addition, comparative phylogeography (the phylogeography of multiple co-distributed taxonomic taxa) helps to reveal previously unrecognised biogeographic patterns and areas (Moritz & Faith, 1998; Arbogast & Kenagy, 2001) which may have conservation values.

Genome size

Genome size refers to the total amount of DNA per nucleus in a species. The term is used to describe holoploid genome size and monoploid genome size with the former referring to the total amount of DNA, irrespective of ploidy level, whereas the latter refers to the genome size of one set of chromosomes. Holoploid genome size is abbreviated to "C-value" and monoploid genome size to "Cx-value". For example 1C-value and 2C-value mean the genome size of unreplicated gametic and somatic DNA amount, respectively. The 1Cx-value equals 1C-value in diploids whereas in polyploids it is the value of 2C-value divided by the ploidy level, and is smaller than the 1C-value. Genome size is measured in picograms (pg) or as the number of nucleotide base pairs in megabases (Mb) (1pg = 978Mb). The fact that genome size is not proportional to the developmental complexity of an organism is referred to as the 'C-value paradox' or 'C-value enigma' (Gregory, 2001).

In angiosperms, genome size has a 2400-fold variation, with a majority of species having a small genome size (Pellicer *et al.*, 2010; Leitch & Leitch, 2012). Natural selection may favour small genome size with large genome size being constrained (Knight *et al.*, 2005). For example, a recent study has shown that plants with larger genome size demand a higher-level of nitrogen (N) and phosphate (P) than plants with smaller genome size (Šmarda *et al.*, 2013). Also, plants with smaller genome size can be more invasive possibly due to their ability to exhibit higher vegetative growth (Suda *et al.*, 2015). In addition, temperature can also influence genome size (Grime & Mowforth, 1982).

Genome size variation is commonly observed. The underling mechanisms of genome size variation may involve differences in the degree of transposable element amplification, ancestral polyploidy (Bennetzen *et al.*, 2005) or infrequent DNA removal (Kelly *et al.*, 2015).

Introduction to relevant methodologies

Molecular phylogenetics

There are two broad categories of phylogenetic tree building methods: distance methods and character-based methods, with the former including the neighbour joining method (NJ) and unweighed pair group method with arithmetic means (UPGMA), while the latter includes maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI). Distance methods convert aligned sequences into a matrix of pairwise distances between sequences and use this matrix to compute branch lengths and branch orders. Character-based methods analyse each site within the multiple alignments. MP seeks the tree with the minimum number of changes assuming that the most likely scenario is the one with the fewest series of events. ML and BI are model-based methods with the former searching the tree that maximises the probability of observing the data given that tree (Felsenstein, 1981) whereas the latter looks for multiple trees of roughly equal likelihood (Hall, 2004).

Commonly used loci for phylogenetic studies in plants are the internal transcribed spacers of ribosomal DNA: internal transcribed spacer 1 (*ITS1*), *5.8S* and internal transcribed spacer 2 (*ITS2*). Its multiple copies in the genome make it easy to amplify; the universal primers for many plant taxa and the intra-genomic uniformity of these copies due to concerted evolution make it suitable for interspecific delimitation (Álvarez & Wendel, 2003). Concerted evolution is a process in which multiple copies of a region are homogenised due to mechanisms such as unequal crossing or gene conversion. Concerted evolution can happen toward either of the parental type. Inter-specific relationships can be assessed if complete concerted evolution of ITS occurs across species. In a recently-formed hybrid, there are two types of ITS present corresponding to each of its parental species. Hence, the parental origins of a hybrid can be traced by comparing their ITS sequences. However, if the ITS sequences of the hybrid are homogenised toward one parent, the information on the other parent will be lost (Álvarez & Wendel, 2003).

Construction of a phylogeny based on multiple loci has become increasingly common as next generation sequencing (NGS) can cheaply sequence many loci across genomes. Many studies have shown that phylogenetic trees based on different loci often have conflicting topology due to various factors such as ILS and introgressive hybridisation

(Degnan & Rosenberg, 2009). Hence, computing a consensus among gene trees is not simple. Several coalescence-based methods have been devised for estimating the species tree, such as the Maximum Pseudo-likelihood Estimation of Species Trees (MP-EST) method (Liu *et al.*, 2010) and the Species Tree estimation using Average Ranks of coalescence (STAR) method (Liu *et al.*, 2009a). MP-EST uses the frequencies of gene trees of triplets of taxa to estimate the topology and branch lengths (in coalescent units) of the overall species tree, whereas STAR computes the topological distances among pairs of taxa as the average of the ranks (number of nodes toward the root node) of those taxon pairs across nodes in the collected gene trees.

Flow cytometry

Flow cytometry is a method widely employed in plant genome size estimation (Doležel & Bartos, 2005; Bennett & Leitch, 2011). It can be applied to identify hybrids between species of differing ploidy level and to distinguish morphologically similar species that differ in ploidy level. It generally involves preparation of nuclei suspensions, fluorescent staining of nuclear DNA and inference of genome size based on a standard of with known genome size (Doležel & Bartos, 2005; Doležel *et al.*, 2007). The nuclei of the species and the standard can be stained with propidium iodide (a DNA intercalator) or with DAPI (binding preferentially to AT-rich regions). Multiple buffers have been used to suspend the nuclei for different plant species to deal with the diversity of chemicals contained across species. A laser-based instrument is commonly used for flow cytometry. Single-cell particles, attached to fluorescent probe, are struck by the laser beam when they pass in single file through a liquid stream. This yields information about the particles based on light-scattering and fluorescence data, which can be analysed using the software associated with the cytometer (Jaroszeski & Radcliff, 1999).

Microsatellite markers

Microsatellites are tandem repeats of units of one to six nucleotides found in the nuclear or plastid genome of many taxa, of which dinucleotide repeats are most commonly used in research. Microsatellites are also referred to as simple sequence repeats (SSR), short tandem repeats (STR) and variable number tandem repeats (VNTR). Microsatellite markers have several advantages over other types of molecular markers (Sunnucks, 2000) such as a high mutation rate (Ellegren, 2000) and a short typical length of 100-300bp, which makes degraded DNA useable for analysis. In addition, SSRs can be multiplexed by labelling primers with different fluorescences

(Guichoux *et al.*, 2011) or the same fluorescence if there is a distinguishable length difference in PCR amplicons. Multiplexing several microsatellites can reduce the cost and increase genotyping accuracy. Due to its advantages, microsatellite genotyping is widely used to study hybridisation and introgression (Randi, 2008; Trigo *et al.*, 2013; McIntosh *et al.*, 2014), ILS (Edwards *et al.*, 2008), hybrid zones (Lexer *et al.*, 2007), parentage analysis (Jones & Ardren, 2003), population demography (Sakaguchi *et al.*, 2013) and population genetic structures (Haasl & Payseur, 2011; Barriball *et al.*, 2015). Despite these advantages, there are some potential problems using microsatellites (Hoffman & Amos, 2005), such as large allele dropout (Miller *et al.*, 2002; Johnson & Haydon, 2006), null alleles (Callen *et al.*, 1993; Pemberton *et al.*, 1995; Dakin & Avise, 2004), homoplasy (Grimaldi & Crouau-Roy, 1997; Estoup *et al.*, 2002) and unclear mutational mechanisms (Ellegren, 2004; Selkoe & Toonen, 2006). Microsatellites can be obtained from those developed for closely-related species or be designed *de novo*.

Analysing microsatellite data for diploid organisms is much easier than for polyploids, because for the latter it is difficult to infer the allele frequencies from heterozygous phenotypes. For example, in a tetraploid, there are different genotypes for AB in which A and B denotes two alleles, such as AAAB, AABB and ABBB. A few methods have been proposed based on signal intensity for each allele (van Dijk *et al.*, 2012; Cuenca *et al.*, 2013). If allele A and B has similar signal intensity in a tetraploid, the genotype is probably AABB and if the signal intensity of A is threefold that of B, it is likely to be AAAB. However, other factors can influence the signal intensity, such as allele size. Currently, one way around this problem is to transform allele presence and absence into binary data prior to subsequent analysis (Sampson & Byrne, 2012); this can be done in the R package POLYSAT (Clark & Jasieniuk, 2011).

Restriction site associated DNA sequencing (RADSeq)

NGS technologies have revolutionised biological research, enabling millions of DNA reads to be obtained within a short period of time and at a low price per base. Restriction site associated DNA sequencing (RADSeq) involves a special library preparation method for high throughput sequencing-by-synthesis technologies, which has been used for various purposes (Etter *et al.*, 2011; Wang *et al.*, 2013). Millions of the flanking regions of restriction cutting sites are sequenced on Illumina platforms. Several protocols for RAD library preparation have been proposed, namely the original RAD protocol (Miller *et al.*, 2007; Baird *et al.*, 2008), double digest RAD (Peterson *et*

al., 2012), ezRAD (Toonen *et al.*, 2013) and 2bRAD (Wang *et al.*, 2012). Each protocol has its own advantages and disadvantages (Puritz *et al.*, 2014). As reagents can be easily obtained, many researchers prepare RAD libraries on their own; this may save much time compared with waiting for results from sequencing centres.

Introduction to the study system: *Betula* (Betulaceae)

General overview of the genus Betula

Betula is a woody angiosperm genus which together with *Alnus*, *Carpinus*, *Corylus*, *Ostrya* and *Ostryopsis* comprises the family Betulaceae. *Alnus* is sister to *Betula* (Chen *et al.*, 1999; Ashburner & McAllister, 2013). *Betula* includes ~60 species, subspecies or varieties with ranges across the northern hemisphere and is of great ecological and economic value. Some species are widely used in horticulture and forestry, such as *B. pendula* and *B. utilis*, whereas other species such as *B. alnoides* and *B. maximowicziana* are important trees for wood production (Ashburner & McAllister, 2013). Despite the fact that some species are widespread, some have very narrow distributions and are listed as endangered in the IUCN Red List, such as *B. calcicola*, *B. corylifolia* and *B. globispica* (Shaw *et al.*, 2014). Polyploidy is common within *Betula* with the ploidy level ranging from diploid to dodecaploid and with the corresponding chromosome number from 2n = 2x = 28 to 2n = 12x = 168. Some species have been reported to contain more than one cytotype, such as *B. chinensis* (6x and 8x), and *B. dahurica* (6x and 8x) (Ashburner & McAllister, 2013).

Classification of *Betula* is difficult with several classifications proposed, i.e., by Regel (1865), Winkler (1904), De Jong (1993) and Skvortsov (2002). In a recent monograph, Ashburner and McAllister classified *Betula* into four subgenera and eight sections: subgenera *Acuminata* (section *Acuminatae*), *Aspera* (sections *Asperae* and *Lentae*), *Betula* (sections *Apterocaryon*, *Betula*, *Costatae* and *Dahuricae*) and *Nipponobetula* (section *Nipponobetula*). This classification is based on Skvortsov (2002) but placed *Acuminata* as a subgenus rather than a section of subgenus *Betula*. So far, this is the most comprehensive monograph of *Betula* and gives a detailed account of its cultivation, biogeography and classification. In addition, it incorporates the chromosome number of nearly all described species and describes the morphology of each species based not only on specimens but also on living trees from botanic gardens.

In the past decade, molecular phylogenies of *Betula* species have been constructed to evaluate the previously proposed classifications (Regel, 1865; Winkler, 1904; De Jong, 1993; Skvortsov, 2002). These phylogenies are partially inconsistent and are all partially contradictory to the previously proposed classifications. This discord is usually attributed to introgressive hybridisation and the occurrences of allopolyploidy,

which commonly occur within *Betula* (Anamthawat-Jónsson & Thórsson, 2003; Nagamitsu *et al.*, 2006).

Betula *species in Britain*

In Britain, there are three *Betula* species: diploid *B. nana* (2n = 2x = 28), diploid *B. pendula* (2n = 2x = 28) and tetraploid *B. pubescens* (2n = 4x = 56), with the former of section *Apterocaryon* (subgenus *Betula*) and the latter two of section *Betula* (subgenus *Betula*) (Ashburner & McAllister, 2013). *Betula nana* (dwarf birch), a shrub species, is nationally scarce and restricted to the Scottish Highlands with fragmented distributions. It is under active conservation by organizations such as Trees for Life. *Betula pubescens* (downy birch) and *B. pendula* (silver birch) are widespread in Britain with the latter species adapted to drier and warmer habitats than the former. Hence, *B. pubescens* is more concentrated in northern and western parts of Britain whereas *B. pendula* is more common in south and east (Gimingham, 1984). *Betula pubescens* and *B. pendula* are hard to distinguish morphologically as there is a continuum of leaf variation between them (Brown & Tuley, 1971; Atkinson & Codling, 1986). The two species were initially treated as *B. alba* and were later split due to a difference in ploidy level. Atkinson discriminant function (ADF) has been devised to distinguish *B. pubescens* from *B. pendula* based on three leaf characters: Leaf Tooth Factor (LTF, the number of teeth projecting beyond the line connecting the tips of the main teeth at the ends of the third and fourth lateral veins, subtracted from the total number of teeth between these two main teeth), the Distance from the petiole to the First Tooth on the leaf base [in millimetres] (DFT) and Leaf Tip Width (LTF, one quarter of the distance between the apex and the leaf base in millimetres). Hybridisation among these species has been documented and hybrids are thought to occur in many areas in the British Isles (Stace, 2010). Hybridisation may threaten the reproduction of dwarf birch populations and may also generate recombinant genotypes of stress-tolerant tree in the subarctic (Vaarama & Valanne, 1973; Wilsey & Saloniemi, 1999; Karlsson *et al.*, 2000). In Iceland, first-generation hybrids between *B. nana* and *B. pubescens* produce viable pollen, which can back-cross with parental species (Anamthawat-Jónsson & Tómasson, 1990). Introgression has been shown to occur between *B. nana* and *B. pubescens* using cytogenetics (Anamthawat-Jónsson & Thórsson, 2003), morphology (Elkington, 1968; Thórsson *et al.*, 2007) and genetic markers (Thórsson *et al.*, 2001; Palmé *et al.*, 2004) and can enable *B. pubescens* to colonise novel habitats (Eidesen *et al.*, 2015). In Iceland, plants in hybrid zones between *B. nana* and *B. pubescens* are

strictly diploid, triploid or tetraploid at the cytological level, but morphological and genetic intermediates are found in all three of these ploidal levels (Anamthawat-Jónsson & Tómasson, 1990; Thórsson *et al.*, 2001; Anamthawat-Jónsson & Thórsson, 2003; Thórsson *et al.*, 2007). Putative hybrids between *B. nana* and *B. pubescens* have been reported in several locations in Scotland, where they are known as *B. x intermedia* (Kenworthy *et al.*, 1972). Bidirectional gene flow has occurred between *B. pendula* and *B. pubescens*, in Scandinavia and western Russia, but with a bias towards gene flow from *B. pendula* to *B. pubescens* (Palmé *et al.*, 2004), perhaps because gene flow is easier from a diploid to a tetraploid than vice versa (Stebbins, 1971).

Aims and scopes of the thesis

In this PhD thesis, I aim to examine the hybridisation of three *Betula* species in Britain: *B. nana*, *B. pubescens* and *B. pendula*, to evaluate the recent classification of *Betula* and to sequence *Betula* species using RADSeq.

In Chapter 2, I used microsatellites and fossil records to address the following questions: (1) What are the introgression patterns among *B. nana*, *B. pubescens* and *B. pendula*? (2) Is hybridisation a potential threat to *B. nana*? (3) Does climate change threaten *B. nana*? A north-to-south cline of introgression from *B. nana* into *B. pubescens* was observed, suggesting that the rarity of *B. nana* was partly caused by hybridisation with *B. pubescens*. In Chapter 3, I evaluated the Atkinson discriminant function (ADF), a method for distinguishing between *B. pendula* and *B. pubescens*, based on the microsatellite data from Chapter 2. The main finding is that the success rate can be raised to 97.5% by using an ADF of -2 rather than zero as the boundary line between the species. In Chapter 4, I evaluated the classification of *Betula* proposed by Ashburner and McAllister (2013) using ITS sequences. Most samples I obtained have been verified by Hugh McAllister and the genome size of these samples was estimated. The results indicate that specimen misidentifications, hybridisation and introgression, the occurrence of polyploid species and morphological convergence potentially cause the discordance between the ITS tree and the classification of Ashburner and McAllister (2013). In addition, the result shows that high ploidy level birches tend to have narrow distributions and the underling reasons merit further research. In Chapter 5, I developed RAD markers for *Betula* and used them for an initial phylogenomic analysis of diploid species using summary-statistics based methods (STAR and MP-EST), a concatenation method and phylogenetic analysis of the binary presence/absence of RAD loci. All these methods yeilded highly supported

phylogenetic trees. Based on these results, I classified *Betula* into four subgenera and seven sections.

Chapter 2 Molecular footprints of the Holocene retreat of dwarf birch in Britain

Publication information:

This chapter is based on a paper published in *Molecular Ecology*, for which I was the lead author. James Borrell performed ecological niche modelling (ENM) analysis and Richard Nichols helped to do cline analysis using mixed-effect model. All authors on this paper contributed to editing and commenting on the original manuscript.

Wang N, Borrell JS, Bodles WJA, Kuttapitiya A, Nichols RA, Buggs RJA*. 2014. Molecular footprints of the Holocene retreat of dwarf birch in Britain. *Molecular Ecology* 23: 2771-2782.

Summary

Past reproductive interactions among incompletely isolated species may leave behind a trail of introgressed alleles, shedding light on historical range movements. *Betula pubescens* is a widespread native tetraploid tree species in Britain, occupying habitats intermediate to those of its native diploid relatives, *B. pendula* and *B. nana*. Genotyping 1134 trees from the three species at 12 microsatellite loci, we found evidence of introgression from both diploid species into *B. pubescens*, despite the ploidy difference. Surprisingly, introgression from *B. nana*, a dwarf species whose present range is highly restricted in northern, high-altitude peat bogs, was greater than introgression from *B. pendula*, which is morphologically similar to *B. pubescens* and has a substantially overlapping range. A cline of introgression from *B. nana* was found extending into *B. pubescens* populations far to the south of the current *B. nana* range. We suggest that this genetic pattern is a footprint of a historical decline and/or northwards shift in the range of *B. nana* populations due to climate warming in the Holocene. This is consistent with pollen records that show a broader, more southerly distribution of *B. nana* in the past. Ecological niche modelling predicts that *B. nana* is adapted to a larger range than it currently occupies, suggesting additional factors such as grazing and hybridisation may have exacerbated its decline. We found very little introgression between *B. nana* and *B. pendula*, despite both being diploid, perhaps because their distributions in the past have rarely overlapped. Future conservation of *B. nana* may partly depend on minimization of hybridisation with *B. pubescens*, and avoidance of planting *B. pendula* near *B. nana* populations.

Introduction

Patterns of genetic variation within and among present day species provide evidence about past population dynamics and demographics. However, interpretation of such genetic evidence is difficult, with multiple historical scenarios potentially explaining the same data. A recent example is the observation of Neanderthal-like genetic variants in modern human population of Eurasia. This observation has been variously explained by: a single hybridisation event (Green *et al.*, 2010), ancient population structure (Durand *et al.*, 2011; Sankararaman *et al.*, 2012; Yang *et al.*, 2012), or hybridisation at a moving front as modern humans invaded Eurasia (Currat & Excoffier, 2011). Such ambiguous situations may be to some extent resolved by additional data sources such as other genetic markers, sample areas, taxa or fossils (Wall *et al.*, 2013). Multiple data sets from exemplar case studies may aid the interpretation of other systems where only a single set of genetic data is available (Buggs, 2007).

One major historical influence on patterns of extant genetic variation is past climate change. Gradients of genetic diversity within species in temperate regions, and correlation of gene phylogenies with geography, can be interpreted as legacies of post-glacial recolonisation with climate warming (Hewitt, 1999; Avise, 2000; Petit *et al.*, 2003). More detailed evidence about species range shifts in response to climate change may be provided by patterns of genetic exchange between closely related species within hybrid zones (Buggs, 2007): specifically, neutral alleles are expected to introgress from a retreating species into an expanding species, leaving behind a molecular footprint of hybrid zone movement (Buggs, 2007; Currat *et al.*, 2008; Scriber, 2011). Whilst this is a potentially sensitive way of tracing past range shifts, genetic patterns alone may not be sufficient to draw firm conclusions, as illustrated by the case of Neanderthals and modern humans (see above).

Many tree species hybridise extensively with local relatives, making them good study systems for examining patterns of introgression as a consequence of climate change (Petit *et al.*, 1997). There is much evidence that tree species have shifted their latitudinal and altitudinal ranges in response to climate change (Davis & Shaw, 2001), and this process is ongoing as the climate warms (Chen *et al.*, 2011). Evidence for this comes from pollen records (Huntley & Birks, 1983), population genetic variability (Petit *et al.*, 2003) and phylogenies (Himes *et al.*, 2008). In areas bounded by inhospitable habitat, some tree species can only respond to climate change by

contracting, rather than shifting their ranges leading to the possibility of local extinction (Zhu *et al.*, 2012).

In this study, we set out to test the hypothesis that the decline of a cold-adapted tree species during Holocene climate warming in Britain could be traced in patterns of introgression of its alleles into a closely related tree species that is currently widespread. To aid the interpretation of introgression patterns, we also analyse patterns of introgression between the widespread species and another close relative with which it is commonly sympatric. We choose a study system with a good fossil record, a well-characterised ecology, and evidence for frequent hybridisation. This system is the *Betula* species of Britain. The genus *Betula* (birches) consists of wind-pollinated tree species, which frequently hybridise (Nagamitsu *et al.*, 2006; Pálsson *et al.*, 2010).

In Britain, there are three native *Betula* tree species: tetraploid *B. pubescens* and diploids *B. pendula* and *B. nana*. *Betula pubescens* (downy birch) and *B. pendula* (silver birch) are common, widespread and often sympatric or parapatric, with the former adapted to wetter and colder habitats than the latter (Atkinson, 1992). *Betula nana* (dwarf birch) is up to 1m high, widespread in subarctic tundra and subalpine areas (DeGroot *et al.*, 1997), but nationally scarce in Britain and mainly restricted to the Scottish Highlands in fragmented populations (Aston, 1984). It is under active conservation management by organisations such as Trees for Life and Highland Birchwoods. Hybrids between *B. nana* and *B. pubescens* have been recorded in the British Isles (Kenworthy *et al.*, 1972; Crawford, 2008; Stace, 2010). In Iceland, such hybrids have been shown using flow cytometry (Anamthawat-Jónsson *et al.*, 2010), morphology (Elkington, 1968; Thórsson *et al.*, 2007), cytogenetics (Anamthawat-Jónsson & Thórsson, 2003) and genetic markers (Thórsson *et al.*, 2001; Palmé *et al.*, 2004). Morphometric analysis of the preserved ancient pollen of these species, suggested that hybridisation has taken place throughout the Holocene (Blackburn, 1952; Caseldine, 2001).

Betula pubescens is more concentrated in northern and western parts of Britain whereas *B. pendula* is more common in south and east (Gimingham, 1984). The two species are hard to distinguish morphologically as there is a continuum of variation between them (Brown & Tuley, 1971; Atkinson & Codling, 1986). Initially, both were treated as *B. alba* (Linnaeus, 1753) and were split later partly due to the difference in ploidy level (Brown & Aldawoody, 1979; Gill & Davy, 1983; Brown & Williams, 1984). Hybrids between the two are thought to occur in many areas in the British Isles,

some of which are fully fertile (Stace, 2010). Bidirectional gene flow has occurred between *B. pendula* and *B. pubescens*, in Scandinavia and western Russia, but with a bias towards gene flow from *B. pendula* to *B. pubescens* (Palmé *et al.*, 2004), perhaps because gene flow is easier from a diploid to a tetraploid than vice versa (Stebbins, 1971).

We tested the hypothesis that decline of the range of *B. nana* in Britain is evidenced by introgression of *B. nana* alleles into *B. pubescens* populations. We surveyed genetic variation at 12 microsatellite loci in 78 populations of *B. pubescens* and 10 populations of *B. nana* in Britain. We expected overall rates of introgression to be low, due to the ploidy difference between the two species, which should result in at least partial reproductive isolation. As a point of comparison in interpreting our results, we also examined gene flow from diploid *B. pendula* into *B. pubescens*. As this tree is morphologically similar and broadly sympatric with *B. pubescens* we would expect more gene flow to have occurred between these two species. We also use ecological niche modelling (ENM) and pollen records to infer the current potential distribution of *B. nana* and its historical distribution range.

Materials and Methods

Sampling and morphological identification

Leaf and twig samples were collected from naturally occurring *Betula* populations across Britain between April 2010 and August 2013. Samples were pressed and dried in a plant press. Species were identified based on leaf morphology according to the standard guide for UK birch identification (Rich & Jermy, 1998), including the Atkinson discriminant function to seek to distinguish between *B. pendula* and *B. pubescens* (Atkinson & Codling, 1986). In total, 1,134 *Betula* samples were collected from 120 populations (Table 2.1). Of these, 120 samples were provisionally identified as *B. nana*, 169 as *B. pendula* and 845 as *B. pubescens* (including some of possible hybrid origin). Three known F_1 hybrid individuals were also examined, two *B. nana* x *B. pubescens* and one *B. nana* x *B. pendula* which had been grown from seed at Queen Mary University of London.

Microsatellite genotyping

Genomic DNA was isolated from dried cambial tissue or leaves following a modified cetyltrimethylammonium bromide (CTAB) protocol (Wang *et al.*, 2013). The isolated DNA was assessed with a Nanovue spectrophotometer (GE Healthcare, UK) and a 1.0% agarose gel. The DNA was diluted to a final concentration of 5-20 ng/µl for subsequent use. A subset of microsatellite loci developed for *B. pendula* (Kulju *et al.*, 2004) and *B. pubescens* ssp. *tortuosa* (Truong *et al.*, 2005) were used. The 5' terminal of forward primers was labelled with FAM, HEX or TAM. Multiplex PCR reactions were conducted combining four pairs of microsatellites in each multiplex (Table 2.2). In each multiplex reaction, two loci with a significant length difference were labelled using the same dye. The final reaction volume was 7.5 µl, including 3.75 µl QIAGEN Multiplex PCR Master Mix, 0.15 µl of primers (10 µM each in initial volume), 1.55 µl H_2O and 5-20 ng of DNA dissolved in 1.0 µl TE buffer. Two touchdown PCR programs (Mellersh & Sampson, 1993) were used with differing annealing temperatures according to the primers within each multiplex. For Multiplex 1 and Multiplex 2 (Table 2.2), an initial denaturation step at 95°C for 15 mins was followed by 28 cycles of denaturation (94°C for 30 s), annealing (65°C to 62 °C for 90 s) and extension (72°C for 60 s) steps, and a final extension step at 60°C for 30 mins. For Multiplex 3 and Multiplex 4 (Table 2.2), the annealing temperature was

from 62°C to 48°C, with the remaining steps unchanged. Fragment lengths were determined by capillary gel electrophoresis with capillary sequencer ABI 3730xl (Applied Biosystems). To check the reproducibility of our microsatellite analyses, we selected a subset of 26 individuals, and repeated the microsatellite analyses of these for each individual. The results indicated 100% match in the results, suggesting that our microsatellite analyses are highly reproducible. Alleles were scored using the software GeneMarker 2.4.0 (Softgenetics) and checked manually.

Three loci with variable flanking regions were genotyped with two sets of primers each to avoid null alleles. One locus, L52, was discarded due to difficulty in reading alleles. Thus, a total of 12 loci were genotyped in our samples. Individuals with more than two missing loci were excluded, resulting in 1,134 individuals in the final dataset.

Microsatellite data analysis

Principal coordinates (PCO) analysis of microsatellite data was performed using POLYSAT (Clark & Jasieniuk, 2011) implemented in R 2.15.3 (R Develoment Core Team, 2012), based on pair-wise genetic distance calculated by Bruvo's methods (Bruvo *et al.*, 2004). POLYSAT is designed to analyse polyploid microsatellite data by assuming that the allele copy number is always ambiguous in any heterozygotes. POLYSAT was also used to transform the

multilocus allele phenotype for each individual into binary arrays of the presence or absence of each allele for each individual and a further PCO analysis was performed using PAST 1.7.5 (Hammer *et al.*, 2001) using pairwise Euclidean distances (Kloda *et al.*, 2008).

We also analysed the microsatellite data with a Bayesian clustering approach in STRUCTURE 2.3.4 (Pritchard *et al.*, 2000) to identify the most likely number of genetic clusters (K), to complement the inference of three disjunct clusters from PCO analysis and taxonomic classification. This implements algorithms accounting for genotypic uncertainty arising from copy number variation when the data include polyploid cytotypes. Individuals are assigned to genetic clusters based on multilocus genotypes. Putative hybrids and admixed individuals could be identified since they have fractions of genomes from different genetic clusters. We performed ten replicates (1,000,000 generations and a burn-in of 100,000 for each run) at each value of K from one to five under the admixture model with the assumption of correlated allele frequencies among populations. Individuals were assigned to clusters based on the highest membership coefficient averaged over the ten independent runs. The ΔK was calculated based on the rate of change in the log probability of the data between successive K values (Evanno *et al.*, 2005). Replicate runs were grouped based on a symmetric similarity coefficient of >0.9 using the Greedy algorithm in CLUMPP (Jakobsson & Rosenberg, 2007) and visualized in DISTRUCT 1.1 (Rosenberg, 2004). We chose the optimal value of K based on the PCO analysis and the ΔK analysis of the STRUCTURE outputs.

The slopes of the latitudinal clines in the admixture proportions (the STRUCTURE values, logit transformed) were estimated using a mixed effects model, with slope as a fixed effect and population modelled as a random effect, to allow for genetic drift of each population away from the trend. This analysis was implemented using the lme function in R 2.15.3 (Pinheiro & Bates, 2000). Despite logit transformation of the proportions the residuals were slightly asymmetrical so, as an additional test, the null distribution of slopes was estimated by permuting the distance values among populations and repeating the analysis, using a custom script in R 2.15.3.

Population genetic parameters were calculated for the selected 55 populations with at least eight individuals from each population. These include six *B. nana* populations, 39 *B. pubescens* populations and ten *B. pendula* populations. Pair-wise F_{ST} tests based on allele frequency were conducted for these populations in POLYSAT. A matrix of

geographical distance was generated based on latitude and longitude in R package 'fields' (Furrer *et al.*, 2011). A Mantel test with 9,999 permutations was conducted in R package 'ade4' to test for a significant signal of isolation by distance (Dray & Dufour, 2007).

Distribution range modelling

In order to model the potential distribution range of the three *Betula* species in Britain, all available occurrence records for the three species were organised into a single database from a number of sources (Botanical Society of the British Isles, National Biodiversity Network, Highland Birchwoods and Scottish Natural Heritage), resulting in 48,164 records. The data were filtered to include only complete records with a spatial resolution <1km and dated post-1950 in order to remain consistent with available environmental data; this resulted in 11,879 records. Twenty-two bioclimatic variables were considered as possible predictors for *Betula* species distribution. These included 19 bioclimatic variable layers obtained from WorldClim (http://www.worldclim.org) (Hijmans *et al.*, 2005); elevation data, also obtained from WorldClim; and soil type and peat depth (where >2m) variables (categorical) obtained from the European Soil Database v.2, (http://eusoils.jrc.ec.europa.eu). All layers were resampled to 1km resolution and clipped to include only the British Isles using Environmental Systems Research Institute's ArcGIS v.10. Modelling was conducted in Maxent ver. 3.3 (Phillips *et al.*, 2004; Phillips *et al.*, 2006), a maximum entropy based machine learning programme that estimates the probability distribution for species occurrence, based on environmental predictors and presence-only data. We ran Maxent under default settings, with 10 subsampled replicated runs, a limit of 5,000 iterations and 25% of the data partitioned for testing of the model. Maxent was used to calculate the area under the curve (AUC) averaged over the replicate runs, to allow comparison of model performance between the study species. Resulting values range from 0.5 (random) to 1.0 (exact match). The resulting potential species distribution map was then opened and manipulated in ArcGIS. Thresholds probabilities for species presence are unknown, thus the resulting values ranging from 0 to 0.88 and were arbitrarily regrouped into six classes; 0 – 0.15, 0.16 - 0.30, 0.31 - 0.45, 0.46 - 0.60, 0.61 - 0.75 and 0.76 - 0.90.

Table 2.1 Details of populations used in this study.

Species	Latitude	Longitude	Approximate localities
B. nana	58.41939	-4.4163	Ben Loyal
B. nana	57.686981	-4.63329	Ben Wyvis
B. nana	57.228046	-4.74222	nr Alti Bhlaraidh
B. nana	57.226832	-4.75489	nr Lochan a'Chlaidheimh
B. nana	57.226193	-4.82272	nr Loch na Beinne Baine
B. nana	57.065987	-2.94768	Ben Gulabin
B. nana	57.022834	-3.626	Mar Estate ne Braemar
B. nana	56.992377	-3.79476	nr Achlean
B. nana	56.918425	-3.20343	Desside
B. nana	56.839519	-3.46695	Ben Gulabin
B. nana	56.626657	-4.75028	Rannoch Moor
B. pubescens	58.8919	-3.38376	Berriedale Wood, Orkney
B. pubescens	58.528058	-4.20936	Bettyhill
B. pubescens	58.500686	-4.37482	Tongue
B. pubescens	58.485084	-4.66051	Loch Eriboll (E)
B. pubescens	58.484644	-4.22032	The Crawford Population
B. pubescens	58.47371	-4.42447	Tongue
B. pubescens	58.44204	-4.42528	Ben Loyal
B. pubescens	58.422395	-4.99334	Achylyness
B. pubescens	58.25238	-5.02184	Kylesku
B. pubescens	58.032116	-4.41898	Lairg
B. pubescens	57.989316	-5.1135	Drumrunie
B. pubescens	57.876346	-4.35883	Ardgay
B. pubescens	57.799685	-4.24486	btwn Cromarty and Dornoch Firths
B. pubescens	57.762708	-5.03172	Braemore
B. pubescens	57.723204	-4.51912	nr Loch Glass
B. pubescens	57.686806	-4.63329	Ben Wyvis
B. pubescens	57.6779	-4.0049	Cromarty
B. pubescens	57.655448	-4.20962	Newmills
B. pubescens	57.366169	-3.9922	SE of Loch Moy, Highland
B. pubescens	57.323454	-4.44631	Urqhart Castle
B. pubescens	57.276645	-3.5541	Lynemore
B. pubescens	57.221883	-3.30125	Blairnamarrow
B. pubescens	57.118613	-3.90135	S of Aviemore
B. pubescens	57.081654	-3.97945	Kingussie, Highland
B. pubescens	57.100902	-3.1524	S-facing slope above Gairnshiel Lodge
B. pubescens	57.08784	-3.18534	Gairnshiel, Ballater
B. pubescens	57.017476	-3.57134	Mar Estate ne Braemar
B. pubescens	57.000518	-3.07762	Glen Muick
B. pubescens	56.932211	-3.17957	N side of Lock Muick
B. pubescens	56.920811	-3.19743	Loch Muick
B. pubescens	56.839193	-3.46647	Ben Gulabin
B. pubescens	56.757025	-5.18886	Loch Linnhe

B. pubescens	56.687358	-3.40676	nr Blairgowrie
B. pubescens	56.62716	-4.74961	Rannoch Moor
B. pubescens	56.415127	-4.50626	Crianlarich
B. pubescens	56.376311	-4.64257	Crianlarich
B. pubescens	56.280697	-2.89677	Bankhead Moss
B. pubescens	56.252788	-4.28193	Callander
B. pubescens	55.241347	-2.21196	Northumberland
B. pubescens	56.232813	-4.70146	Loch Lomond
B. pubescens	55.226935	-3.42971	Johnstonebridge, Dumfries
B. pubescens	54.833235	-1.90462	ConsettWood
B. pubescens	54.76765	-1.90237	TunstallReservoir
B. pubescens	54.610073	-2.52893	Brampton, Cumbria
B. pubescens	54.508494	-1.10609	RoseberryTopping
B. pubescens	54.394734	-1.20611	GrouseFarm, N Yorks
B. pubescens	54.39239	-2.00341	BirkPark, Yorks
B. pubescens	54.39126	-1.82729	RichmondQuarry, Yorks
B. pubescens	54.008643	-1.38711	Flaxby, North Yorkshire
B. pubescens	53.99877	-1.88831	Bolton Abbey
B. pubescens	53.801452	-2.40882	ScoutCamp, Lancs
B. pubescens	53.4335	-1.9525	GlossopWood, Derby
B. pubescens	52.929519	1.203763	nr Bywater PG
B. pubescens	52.835273	0.99854	nr Tipples Farm
B. pubescens	52.815356	1.045277	Hindolveston Wood
B. pubescens	52.693503	1.46372	Horning Norfolk
B. pubescens	52.493378	0.987717	Quidenham
B. pubescens	52.454126	0.996216	Eccles Carr, Norfolk
B. pubescens	52.348312	-1.4489	RytonWood, Warwick
B. pubescens	51.846921	-3.18155	BreconBeacons2
B. pubescens	51.711681	0.570787	Sporhams Lane
B. pubescens	51.174753	0.337031	Capel, Kent
B. pubescens	51.150878	0.491267	Widehurst Wood
B. pubescens	51.112171	-0.93291	Long Copse
B. pubescens	51.085146	-0.1601	Turners Hill
B. pubescens	51.04168	-0.61157	Ebernoe
B. pubescens	50.977624	-0.58132	Corner Copse
B. pubescens	50.937354	-3.29404	DevonM5/B3391
B. pubescens	50.919858	-0.4758	nr Cootham Pre-school
B. pubescens	50.900505	-1.57963	Lyndhurst
B. pubescens	50.846874	-1.45321	Denny Lodge
B. pubescens	50.817012	-1.53334	nr Roundhill Campsite
B. pubescens	50.520095	-3.80108	DartmoorHotel
B. pubescens	50.516668	-3.80806	DartmoorVerge
B. pubescens	50.498824	-3.79103	DartmoorNTwood
B. pubescens	50.523693	-3.82075	DartmoorBurrator
B. pubescens	50.491992	-4.04484	DartmoorTree2

B. pendula	58.528058	-4.20936	nr Strathnaver museum
B. pendula	57.799685	-4.24486	nr Feith Ruadh
B. pendula	57.678385	-4.00164	nr Sutors of Cromarty
B. pendula	57.655955	-4.20736	Dingwall
B. pendula	57.36857	-4.49608	Inverness
B. pendula	57.323454	-4.44631	nr Urquhart Castle
B. pendula	57.168489	-3.8276	nr Spey Lodge
B. pendula	57.052712	-3.14613	nr Easter Micras Bum
B. pendula	57.000518	-3.07762	nr Alt Dowrie
B. pendula	56.572309	-3.31653	nr Coupar Angus Rd
B. pendula	56.550905	-3.55053	nr Birnam
B. pendula	54.833235	-1.90462	nr South Horseleyhope Bum
B. pendula	54.507909	-1.11099	nr Roseberry Topping
B. pendula	54.008314	-1.38737	nr Harrogate Paintball Center
B. pendula	53.99877	-1.88831	nr Crabtree JC
B. pendula	53.4335	-1.9525	nr Quinlan Autos
B. pendula	53.387802	-1.05901	nr Travelodge hotel
B. pendula	53.162018	-1.62461	nr Stantonlees Chapel
B. pendula	52.929519	1.203763	nr Bywater PG
B. pendula	52.569649	1.007121	nr the Granary Crown Farm
B. pendula	52.564542	-3.15843	nr Jamesford Farm
B. pendula	52.347661	-1.45189	nr Ryton Pools Country Park
B. pendula	52.036003	-2.34518	nr Zephir Lidar
B. pendula	51.915225	-3.1746	nr Cwm-rhos Brook
B. pendula	51.815496	-3.05377	nr Lianwenarith Baptist Church
B. pendula	51.711681	0.570787	nr Fitzwalter Ln
B. pendula	51.174919	0.335751	nr Half Moon Ln
B. pendula	51.04168	-0.61157	nr Simmonds Saws
B. pendula	50.817012	-1.53334	nr Roundhill Campsite
B. pendula	50.521522	-3.80442	nr Chase Hill Lodges
B. pendula	50.523089	-3.82133	Widecombe in the Moor
B. pendula	50.919858	-0.4758	nr Cootham Pre-school

Table 2.2 Details of microsatellite primers used in the present study.

Locus	Dye	Sequences	Allele Size	Repeat	Multiplex
L1.1*	HEX	ACGCTTTCTTGATGTCAGCC TCACCAAGTTCCTGGTGGAT	168–209	$(GA)_4AA(GA)_{10}$	Multiplex 1
L3.1*	FAM	CTCCTTAGCTGGCACGGAC CCCTTCTTCATAAAACCCTCAA	219–241	$(CT)_3CC(CT)_2CC(CT)_{13}AT(CT)_5$	
L13.1	FAM	CACCACCACAACCACCATTA AACACCCTTTGCAACAATGA	93–108	$(CA)_3(GA)_{14}$	
L012	TAM	TGGTTGACGTGACGTTGATT GGCCCATAGGGAAGATAAGC	210–222	$(GA)_6TA(GA)_6$	
L52*	FAM	AGCTACCCCTGGTCCACTTT CCGCCTTGGATTTCACTAAA	250–272	$(CT)_{12}$	Multiplex 2
L5.4*	HEX	AAGGGCACCTGCAGATTAGA AAAATTGCAACAAAACGTGC	230–262	$(TC)_{26}$	
L7.1a	HEX	GTTTTGGGTTTCCACTTCCA ACTGGTAATACCTTTACCAAGCC	146–152	$(CT)_{12}CCTT(CT)_4$	
L7.3	TAM	GGGGATCCAGTAAGCGGTAT CACACGAGAGATAGAGTAACGGAA	178–226	$(GT)_{18}(GA)_{14}$	
L1.1	FAM	TTTCCAACGCTTTCTTGATG TGGATAAGGAAGGGCATGTC	152–206	$(AG)_4AA$	Multiplex 3
L2.3	HEX	CGGGAAGATATGCAGTGTTT TTGGCGGGTGAAGTAGAC	208–252	$(AG)_{16}$	
L3.1	HEX	CACACTGCTGCCTGA TCATAAAACCCTCAAAGAAT	134–166	$(CT)_{13}A(TC)_6$	
L021	TAM	TCTACGCTGTGACCAGTC AGAATCCTAGCCTTTTCAAT	168–236	$(CT)_{14}$	
L2.5	FAM	CTATATTGGCTCCAAGCAC ACACCCACACTGACAGATAA	94–128	$(CT)_9$	Multiplex 4
Bo.F330	FAM	TGGCAGCACGAAAGT TGGGAATGAGAGAACAAG	172–210	$(TC)_{14}$	
Bo.F394	HEX	AATGCAGCATCTCTTACC CACGCAATAATATGGAAA	128–194	$(TC)_{13}$	
L5.4	TAM	GAAAGCATGAGACCCGTCTT AACCTAAACAGCCTGCCAAA	134–188	$(TC)_{26}$	

*: Discarded due to redundancy or difficulty in allele assigning.

Niche overlap between species was measured using Schoener's D (Schoener & Gorman, 1968), and the I statistic (Warren *et al.*, 2008), calculated in ENMTools v.1.4.3 (Warren *et al.*, 2010). Similarly, species range overlap was also tested in ENMTools v1.4.3, over a range of manually defined presence probability thresholds to explore the characteristics of the data. We chose a conservative value of 0.45, though we note that the comparative relationships between the three species remain consistent over a broad range.

Pollen record gathering

In order to build a picture of the past distribution of these species in the UK, we examined pollen records of *Betula species* in the European Pollen Database (EPD, http://www.europeanpollendatabase.net/data/). For some pollen cores, palaeobotanists have identified pollen type to the species level; whereas others are identified at the genus level only. We mapped these pollen sites using coordinates given in the EPD. For eight pollen sites, coordinates are not given in the EPD, so we mapped the sites according to the geographical descriptions given in the original literature.

Results

Microsatellite analysis

Broad characterisation of genetic diversity among the three *Betula* species was conducted with Principal Coordinates (PCO) Analysis. The Bruvo's genetic distances of all 1,134 individuals were calculated and scaled. The first axis separated *B. pendula* from a cluster of *B. pubescens* and *B. nana*, and the second axis separated *B. nana* from *B. pubescens* and *B. pendula*. Thus three distinct clusters corresponded to *B. nana*, *B. pubescens* and *B. pendula* (Fig. 2.1).

Genetic admixture among species within individuals was examined with Bayesian analysis using STRUCTURE under the admixture model. Analysis was conducted assuming three populations (K=3) based on clear clustering in the PCO distribution, corroborated by the ΔK criterion. The estimated admixture between *B. pendula* and *B. nana* was negligible (Fig. 2.2 A, B, D), but admixture was inferred between *B. pubescens* and *B. nana* and also between *B. pubescens* and *B. pendula* despite *B. pubescens* being tetraploid (Fig. 2.2 A). Higher levels of admixture from *B. nana* to *B. pubescens* were found in the north than in the south of Britain. The cline of *B. nana* admixture in *B. pubescens* populations was positively correlated with latitude (Fig. 2.3A, $P = 0.0045$). Conversely, the cline of *B. pendula* admixture in *B. pubescens* was negatively correlated with latitude (Fig. 2.3B, $P = 0.0166$), higher levels of admixture being found in the south than in the north of Britain (Fig. 2.2C).

Significant isolation-with-distance was detected among *B. nana* populations (Mantel test, $r = 0.7035$, $P = 0.0086$) and among *B. pubescens* populations (Mantel test, $r = 0.1384$, $P = 0.0093$) but not among *B. pendula* populations (Mantel test, $r = -0.0418$, $P = 0.5709$). Genetic differentiation between *B. nana* and *B. pendula* was higher than between *B. nana* and *B. pubescens*, and between *B. pubescens* and *B. pendula*. Genetic structure was detected among *B. nana* populations but not among either *B. pubescens* or *B. pendula* populations when the three species were analysed independently with the admixture model (Fig. 2.4).

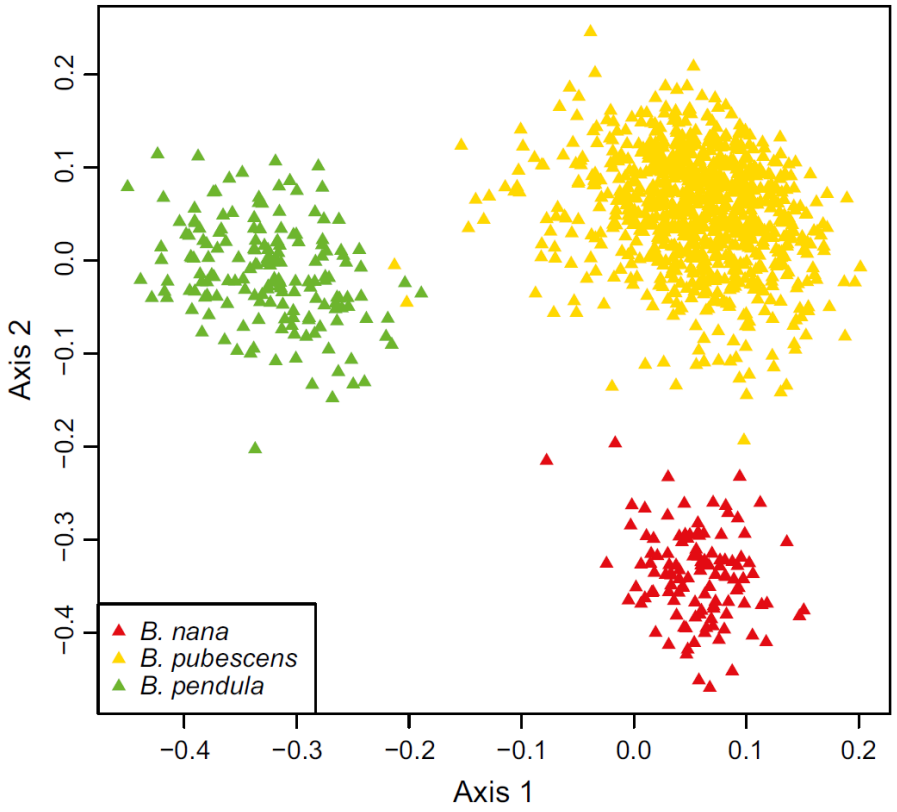


Figure 2.1 Principal coordinate (PCO) analysis of *B. nana*, *B. pubescens* and *B. pendula* based on Bruvo's genetic distance of microsatellite data.

Model-based prediction of past distribution ranges

Ecological niche models constructed with MAXENT from species occurrence records, including herbarium collections, performed well for *B. nana* (AUC = 0.959, s.d. = 0.018) and were satisfactory for *B. pendula* (0.723 ± 0.009) and *B. pubescens* (0.645 ± 0.008). The most important environmental predictors were soil type and annual mean temperature, with the exception of *B. nana* for which altitude was of primary importance. The results suggest that suitable habitats for *B. nana* may currently exist in large areas in the Scottish Highlands, SW England, Wales, middle and North England (Fig. 2.5): an area larger than the area currently occupied by *B. nana*. Suitable habitat for *B. pubescens* and *B. pendula* appears widespread in Britain, the most suitable habitat for *B. pendula* being towards the south and east, and suitable habitat for *B. pubescens* being widespread. Analysis of pairwise niche overlap revealed

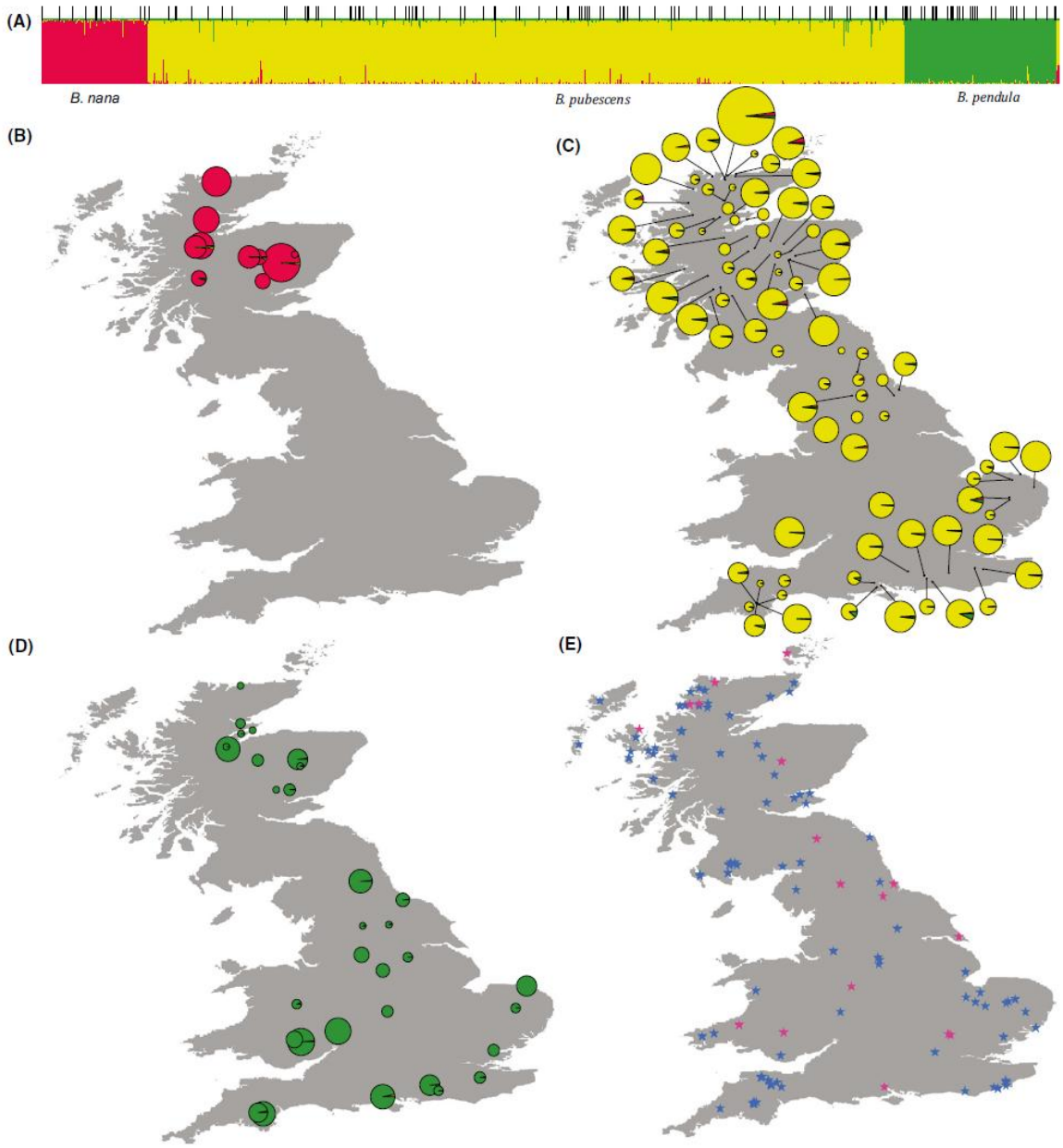


Figure 2.2 Genetic admixture among the three native *Betula* species in Britain, with locations of populations tested, and pollen fossil sites. (A) Sharing of microsatellite alleles among the three species *B. nana*, *B. pubescens* and *B. pendula* shown as a STRUCTURE plot with K = 3 corresponding with the three species. Within each species grouping, populations are ordered by latitude, with more northerly populations to the left hand side. Thin vertical lines above the STRUCTURE plot indicate population divisions. Three known F_1 hybrid seedlings are shown on the far right: *B. nana* x *B. pendula*, *B. nana* x *B. pubescens* and *B. nana* x *B. pubescens*, respectively. (B, C, D) The locations of the sampled populations of *B. nana*, *B. pubescens* and *B. pendula* tested, respectively: pie-charts show the mean proportion of individual genotypes in each population assigned to a particular lineage by STRUCTURE and pie-chart size is proportional to the sample size for each population. The centre of pie-charts represents approximately its sampling locality unless the pie-chart is connected to its sampling locality by a straight line. (E) Pollen sites of *Betula* species across Britain. Red stars represent the pollen sites of *B. nana* and *B.* cf. *nana* and blue stars represent the pollen sites of *Betula* likely to be *B. pubescens* and *B. pendula*.

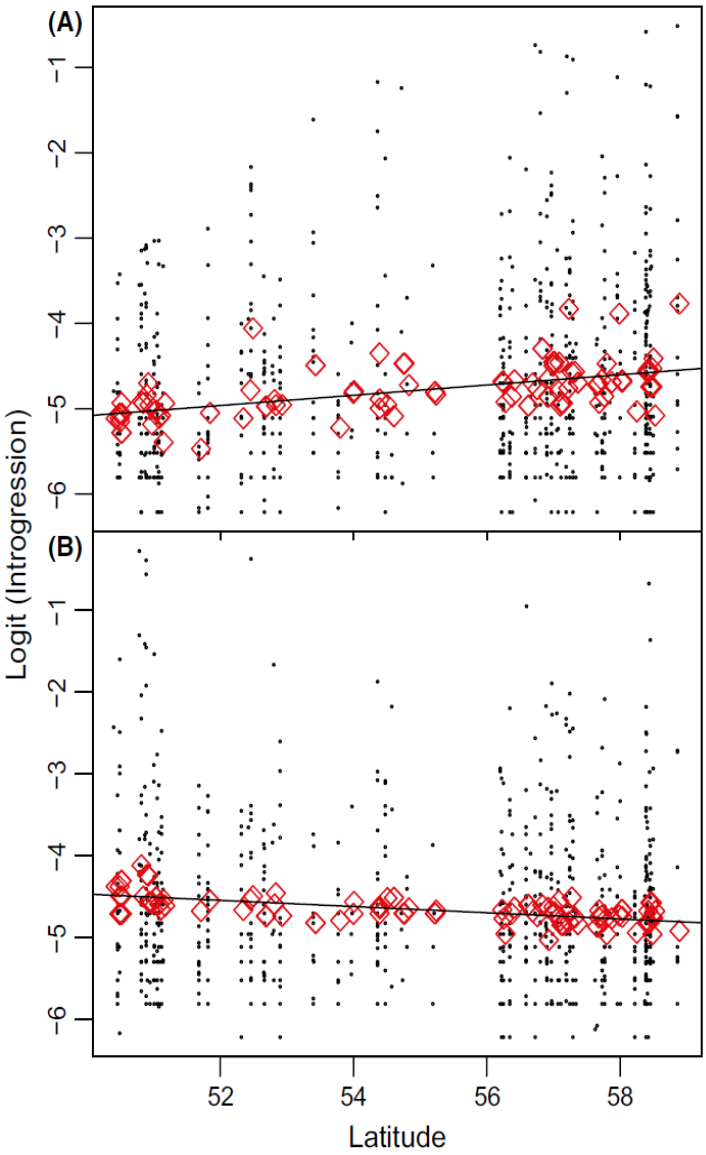


Figure 2.3 Clines of *B. nana* and *B. pendula* admixture into *B. pubescens* populations. The latitude of each sample populations is shown on the horizontal axis, and logit-transformed STRUCTURE admixture proportions for each *B. pubescens* individual are shown as circles. Red diamonds represent the value for each *B. pubescens* population fitted by the mixed effects model. (a) The cline of *B. nana* admixture into *B. pubescens* populations, which showed a significant positive correlation with latitude (P = 0.0045). (b) The cline of *B. pendula* admixture into *B. pubescens* populations, which showed a significant negative correlation with latitude (P = 0. 0166).

considerable similarity between *B. pubescens* and *B. pendula* niches (Schoener's D = 0.82, I = 0.97). There was substantially less overlap when comparing *B. nana* with *B. pubescens* (D = 0.25, I = 0.58) and *B. pendula* (D = 0.18, I = 0.48). Range overlap analysis at a conservative occurrence probability threshold (0.45) identified extensive overlap between *B. pubescens* and *B. pendula* (73%) and small overlap between *B. pubescens* and *B. nana* (5%), but no range overlap between *B. nana* and *B. pendula*.

Suitable habitats for *B. nana* either overlap with or are surrounded by suitable habitats for *B. pubescens* (Fig. 2.5).

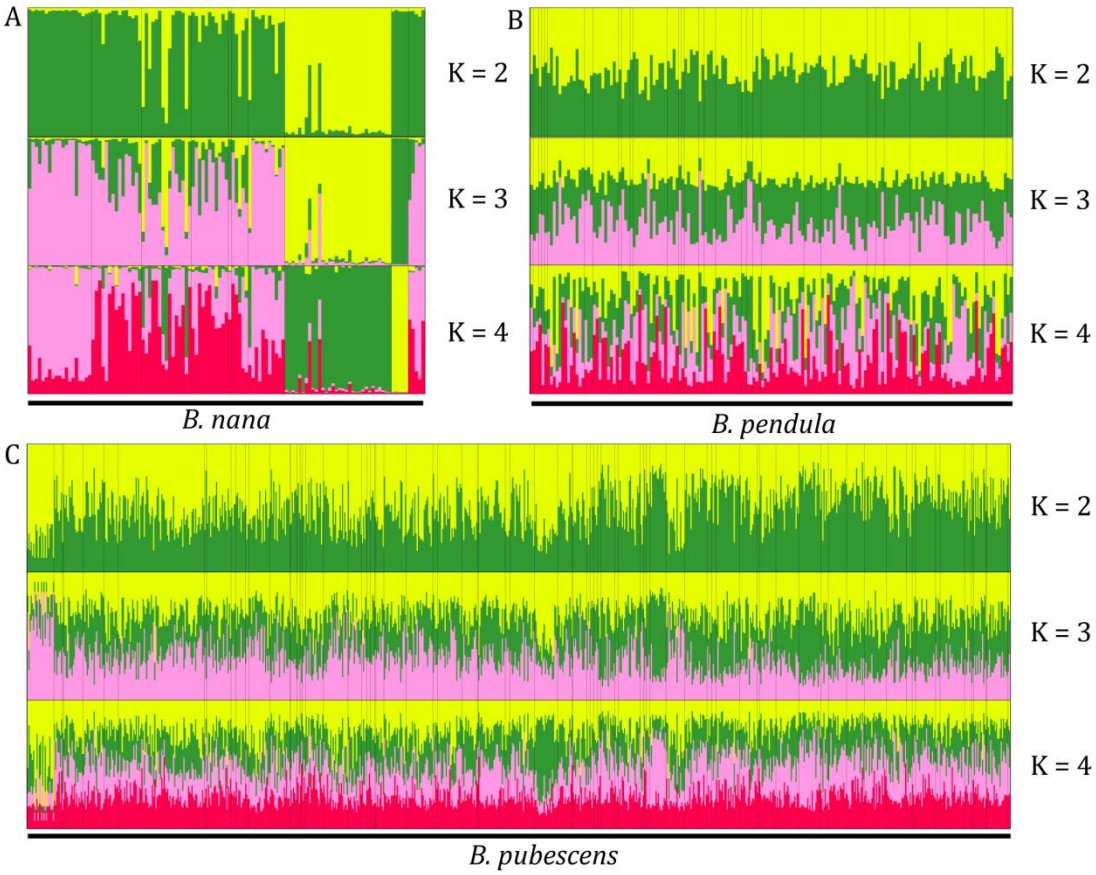


Figure 2.4 The STRUCTURE output of *B. nana*, *B. pubescens* and *B. pendula* separately under the model of admixture, at K = 2, 3 and 4.

Pollen records

Betula ssp. pollen was found recorded for 112 British sites of preserved pollen in the European Pollen Database. The majority most likely represent *B. pubescens* and *B. pendula*, which produce abundant pollen, but 13 sites contained pollen identified as *B. nana* pollen and four contained pollen identified as "*B.* cf *nana*" (Fig. 2.2E). These included sites in the south that are outside the current range of *B. nana* suggesting a much more southerly distribution of *B. nana* in the past. These pollen records provide us with a longer-term view of the past distribution ranges of *B. nana* than herbarium collections.

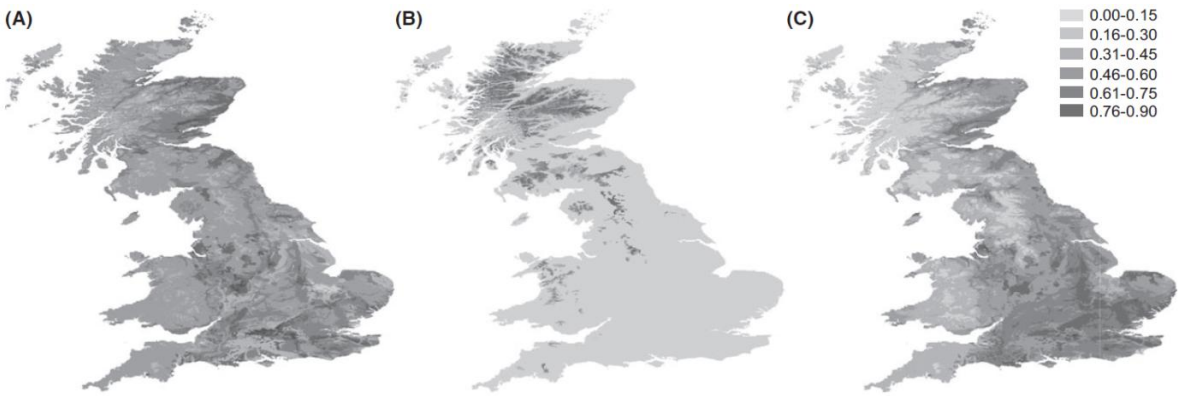


Figure 2.5 Ecological niche model predicted distribution British ranges for (a) *B. pubescens* (b) *B. nana* and (c) *B. pendula*. At an occurrence probability threshold of 0.45 range overlap is as follows: *B. pubescens* and *B. pendula* (73%); *B. pubescens* and *B. nana* (5%); *B. nana* and *B. pendula* (0%).

Discussion

Genetic admixture among closely related species may occur for a variety of reasons: (A) shared alleles may have been inherited from a polymorphic common ancestor due to incomplete lineage sorting, (B) convergent mutations may have caused the same alleles to have arisen independently in different species, (C) alleles may have moved from one species to another via introgressive hybridisation within a framework of stable species' ranges, perhaps assisted by selection, (D) alleles may have moved via introgressive hybridisation with neutral gene flow, increased by the invasion of one species into the range of another. We consider that in the present study the balance of evidence points towards (D), introgressive hybridisation from *B. nana* to *B. pubescens*, due to range expansion of the latter at the expense of the former.

Although incomplete lineage sorting (A) may be frequent among tree species due to their large effective population sizes and long generation times (Bouillé & Bousquet, 2005; Chen *et al.*, 2010), this seems an unlikely explanation for the patterns of allele sharing observed between *B. nana* and *B. pubescens*, because we find a gradient of *B. nana* alleles that increases closer to the current location of *B. nana* populations in the north. If the loci in the study were neutral with respect to selection, which is expected of microsatellite alleles, then incomplete lineage sorting would not be expected to give a geographic signal.

Convergent mutations (B) also seem an unlikely explanation, due to the geographic patterning of *B. nana* alleles in *B. pubescens*, and the neutrality of microsatellite loci. If it were caused by scenario (B), we would also expect symmetric allele sharing between *B. pubescens* and *B. nana*. However, this is contrary to the observed pattern: allele sharing is asymmetric with *B. pubescens* possessing more *B. nana*-type alleles than vice versa (Fig. 2.2A, B, C).

The pattern we observe is therefore likely to be caused by hybridisation. *Betula nana* and *B. pubescens* currently have parapatric distributions and often occur close together in natural environments. Several putative hybrids have been noted by taxonomists in Scotland (Kenworthy *et al.*, 1972) and extensive hybridisation and gene flow have been shown to occur between the two species in Iceland (Anamthawat-Jónsson & Thórsson, 2003; Maliouchenko *et al.*, 2007), Scandinavia and Russia (Maliouchenko *et al.*, 2007). However, the pattern of introgression that we observe is unlikely to have

been caused simply by spread of alleles from the current distribution range of *B. nana*. High genetic differentiation and significant isolation-with-distance among *B. nana* populations suggests that *B. nana* has a low capacity for gene flow, as is to be expected for a shrub producing small amounts of pollen and seed compared to its larger tree relatives (Bradshaw, 1981). Also, because microsatellites markers are expected to be neutral to selection, the presence of *B. nana* alleles in occasional *B. pubescens* populations far from the present range of *B. nana* in the middle of Britain is unlikely to have been caused by natural selection.

The observed level of introgression from *B. nana* to *B. pubescens* is not less than the level of introgression we observed from *B. pendula* to *B. pubescens* (Student's t test, t = 0.082, P = 0.934). This is surprising given that *B. pendula* is a tree that disperses more pollen than *B. nana* and frequently occurs in sympatry with *B. pubescens* in much of its British range (Atkinson, 1992). Given that *B. nana* and *B. pendula* are diploid with the same chromosome number, they are unlikely to differ in chromosomal post-zygotic reproductive isolation with tetraploid *B. pubescens*. Hybrids between *B. pendula* and *B. pubescens* have been recorded in the UK (Brown *et al.*, 1982) and a study of chloroplast introgression in Scandinavia and western Russia found higher rates of introgression between *B. pendula* and *B. pubescens* than between *B. nana* and *B. pubescens* (Palmé *et al.*, 2004). The fact, therefore, that we found similar introgression from *B. nana* to *B. pubescens* and from *B. pendula* to *B. pubescens* requires an explanation.

The most likely explanation of the pattern observed in this study is (D) that we are seeing a trail of introgression resulting from past invasion by *B. pubescens* into the range of *B. nana*. This could explain the high level of introgression found relative to *B. pendula–B. pubescens* introgression and the geographic pattern of introgression observed. This hypothesis fits with the fact that fossils of *B. nana* and *B.* cf. *nana* pollen are distributed across Britain (Fig. 2.2E) showing a larger and more southerly range in the past. Both genetic and fossil evidence therefore point to the northwards movement of the range of *B. pubescens* in the UK, at the expense of *B. nana*, with some hybridisation occurring between them during this expansion/retreat.

What caused this expansion of *B. pubescens* at the expense of *B. nana*? The fact that *B. nana* pollen is found outside the current environmental niche range of *B. nana* suggests that past climate change has played a major role in the species' decline. But the fact that *B. nana* is currently more restricted in its range than the area that it is adapted to

according to the ENM suggests that other factors may also have contributed to the decline of *B. nana*, such as over-grazing by sheep and deer (Tanentzap *et al.*, 2013) and burning of moorland for grouse shooting (DeGroot *et al.*, 1997). A further contributing factor may be pollen swamping of *B. nana* by *B. pubescens*, reducing the production of fertile *B. nana* offspring in *B. nana* populations. We believe that the low levels of introgression found in this study support the pollen-swamping hypothesis. Due to the ploidy difference between *B. nana* and *B. pubescens*, we expect most hybrids to be sterile, so only a minority of hybrids formed will be capable of contributing to introgression between the two species. Therefore the small amount of introgression we observe between *B. nana* and *B. pubescens* suggests that large numbers of hybrids may be forming, as has been found in Icelandic populations of *B. nana* and *B. pubscens* where up to 10% of trees may be hybrids (Anamthawat-Jónsson & Tómasson, 1999; Anamthawat-Jónsson & Thórsson, 2003). Furthermore, the asymmetric pattern of gene flow that we observe suggests that on the rare occasions when hybrids are capable of backcrossing, they do so mainly with *B. pubescens*, rather than *B. nana*. This, and the fact that *B. pubescens* is a tree with far greater pollen dispersal ability than *B. nana*, suggests that *B. nana* ovules may be frequently fertilised by *B. pubescens* pollen. Thus reproduction of *B. nana* may be reduced by the production of (mainly sterile and non-backcrossing) hybrids with *B. pubescens*. Such a dynamic has been shown to occur in a hybrid zone between diploid and hexaploid *Mercurialis annua*, where the hexaploid form is apparently being eliminated by the diploid form due to pollen swamping and the production of sterile hybrids (Buggs & Pannell, 2006). Even when hybrids are not mainly sterile, pollen swamping can still contribute to the advance of one species' range at the expense of another, for example, pollen swamping of *Quercus robur* by *Q. petraea* seems to allow the latter in invading the range of the former (Petit *et al.*, 2004).

We found very little introgression between *B. nana* and *B. pendula*, despite the fact that a reproductive barrier due to ploidy does not separate them as they are both diploids. While we do not know if other reproductive barriers do separate them, we have found diploid hybrids when growing up seeds collected from *B. nana* populations in Scotland, in an area recently planted with *B. pendula* in afforestation, suggesting that *B. nana – B. pendula* hybrids do form in Scotland. The most probable explanation for the lack of introgression between the two species in our study is the disjunct nature of their natural distributions: the environmental niches of the two rarely overlap (Fig.

2.5). Corroborating this finding, a six year study in Sweden showed the germination rates of *B. pendula* seeds to decrease strongly with altitude (Holm, 1994). *Betula nana* is adapted to cold and wet habitats (DeGroot *et al.*, 1997) whereas *B. pendula* prefers warm and dry habitats (Gimingham, 1984). *Betula nana* commonly grows above the treeline whereas *B. pendula* grows in regions with low altitude usually below a few hundred meters (Gimingham, 1984). Maintenance of this geographical separation between *B. nana* and *B pendula* may be key to preventing future hybridisation between them.

We conclude that a balance of evidence from both genetic data and fossils suggests that a zone of hybridisation between *B. nana* and *B. pubescens* moved northwards through the UK since the last glacial maximum, leaving behind a footprint of introgressed genes in the genome of *B. pubescens*. The decline of *B. nana* due to climate change may have been exacerbated by hybridisation with *B. pubescens*. Today, *B. nana* is nationally scarce in Britain and under active conservation management. Successful conservation of *B. nana* may partly depend on minimisation of future gene flow from *B. pubescens*. However, a bigger threat may be hybridisation with *B. pendula*; although there appears to have been little hybridisation between *B. nana* and *B. pendula* in the past, this may be due to ecological separation rather than reproductive incompatibility (Wilsey *et al.*, 1998), and planting of *B. pendula* saplings in areas where *B. pendula* could not establish from seeds may be causing a new anthropogenic threat to the reproduction of *B. nana*.

Chapter 3 Is the Atkinson discriminant function a reliable method for distinguishing between *Betula pendula* and *B. pubescens*?

Publication information:

This chapter is based on a manuscript published in *New Journal of Botany*. All authors contributed to editing, proofreading and commenting on the published manuscript.

Wang N., Borrell JS, Buggs RJA*. 2014. Is the Atkinson discriminant function a reliable method for distinguishing between *Betula pendula* and *B. pubescens*? *New Journal of Botany* 4: 90-94.

Summary

Betula pendula and *B. pubescens* are common tree species of Europe that differ in ploidy level. The continuum of morphological variation between the two species makes them hard to differentiate in the field. The Atkinson Discriminant Function (ADF) based on leaf shape was proposed in 1986 as a metric to distinguish them and has since become a standard approach. Here, we test this method on 944 trees sampled across Britain against species' discriminations made using 12 microsatellite loci. The ADF misidentified six of 780 *B. pubescens* trees and 28 of 164 *B. pendula* trees. This success rate of 96.4% is higher than that found by Atkinson & Codling (1986) based on a smaller sample for which chromosomes had been counted. The success rate can be raised to 97.5% by using an ADF of -2 rather than zero as the boundary line between the species. With this improvement, error rates of over 10% occur for trees with ADF ranging from -11 to +3.

Introduction

Betula pendula Roth. (silver birch) and *B. pubescens* Ehrh. (downy birch) are closely related north-temperate tree species that have considerable morphological similarity but are genetically separate due to a difference in ploidy level (Brown & Tuley, 1971; Atkinson, 1992). *Betula pendula* is a diploid (2n = 28) whereas *B. pubescens* is a tetraploid species (2n = 56). The two species have considerably overlapping ranges though slightly different ecological niches, with *B. pendula* preferring drier and warmer habitats to *B. pubescens* (Atkinson, 1992). In a 1986 *Watsonia* article, M. D. Atkinson & A. N. Codling proposed a discriminant function to distinguish between *B. pendula* and *B. pubescens* based on leaf shape measurements. Atkinson & Codling (1986) compared the Atkinson Discriminant Function (ADF) with chromosome number data for 104 trees, finding that the ADF correctly classified 97 of the trees. All seven misclassified trees were *B. pubescens* mistaken for *B. pendula* (Atkinson & Codling, 1986). The ADF function is included in Stace's (2010) Flora and has become a standard approach to identification for field botanists (Rich & Jermy, 1998).

The morphological continuum between *B. pendula* and *B. pubescens* has been the subject of several studies (Atkinson, 1992; Howland *et al.*, 1995; Franiel & Więski, 2005; Kovacic & Nikolic, 2005), which collectively have centred on two major hypotheses. One hypothesis is that the intermediate forms are hybrids (reviewed in Section VII(b) of Atkinson, 1992), though current evidence suggests that hybrids are rare and of low fertility, and often resemble *B. pubescens* in their morphology (Brown *et al.*, 1982; Atkinson, 1992). The second hypothesis is that intermediate morphologies arise due to phenotypic plasticity in different environments (Gill & Davy, 1983) or ecotypic variation (Pelham *et al.*, 1988). For example, *B. pendula* leaf shape is different between unpolluted and zinc/lead contaminated habitats (Franiel & Więski, 2005) and between bogs versus heaths (Davy & Gill, 1984). However, in Croatia, *B. pendula* leaf character variation was poorly correlated with the environmental gradients (Kovacic & Nikolic, 2005).

A genetic contributor to the morphological continuum between *B. pendula* and *B. pubescens* may be the allopolyploid nature of the *B. pubescens* genome. Though the progenitors of *B. pubescens* have not been confirmed, several researchers have suggested that *B. pendula* may have been involved (Walters, 1968; Howland *et al.*, 1995). This may explain why the two species have proved hard to discriminate with

rDNA repeat length analysis and RAPD data (Howland *et al.*, 1995), and phylogenetic analysis based on AFLP markers (Schenk *et al.*, 2008) and ITS sequence (Li *et al.*, 2005).

In a recent study of genetic variation of *Betula* species in Great Britain, we analysed 12 microsatellite loci in 1,134 trees identified as *B. nana*, *B. pendula* or *B. pubescens* (Wang *et al.*, 2014a) with the STRUCTURE software package. STRUCTURE uses multilocus allele frequencies to assign individuals to populations (Hubisz *et al.*, 2009), and gave clear genetic discrimination between the three species (Wang *et al.*, 2014a). Here, we re-examine the *B. pubescens* and *B. pendula* trees (for which we have suitably preserved leaves) from this study, comparing their ADF scores with their species allocation as determined by microsatellite genotyping. We thereby accurately assess the reliability of the ADF for distinguishing between the two species.

Materials and Methods

Sampling

Leaf and twig samples were collected from naturally occurring *B. pendula* and *B. pubescens* populations across the United Kingdom between April 2010 and August 2013 (for localities and other details see Chapter 2). Samples were pressed and dried. Species were initially identified based on leaf morphology according to the birch entry in Rich & Jermy (1998). In total, 944 *Betula* individuals were collected from 105 populations.

STRUCTURE analysis

We analysed microsatellite data from 12 loci for the 944 trees in STRUCTURE 2.3.4 (Pritchard *et al.*, 2000). We performed 1,000,000 generations and a burn-in of 100,000 as recommended (Gilbert *et al.*, 2012), specifying two populations (K = 2), for each run of three replicates. We used a model in which STRUCTURE estimates the posterior probability that each individual is from each population. Replicated runs were grouped based on the similarity coefficient of >0.9 using the Greedy algorithm in CLUMPP (Jakobsson & Rosenberg, 2007) and visualised in DISTRUCT 1.1 (Rosenberg, 2004). Using the dataset, we rerun STRUCTURE analysis with the same parameters as specified above but used the admixture model which estimates the admixture values from the two populations for each individual. We then plotted the admixture values against the ADF scores to evaluate if there is a relationship between hybridisation and ADF score.

ADF Scoring

We applied the ADF to our samples according to the formula in Atkinson & Codling (1986): ADF score = 12LTF + 2DFT – 2LTW – 23 (where LTF is Leaf Tooth Factor [the number of teeth projecting beyond the line connecting the tips of the main teeth at the ends of the third and fourth lateral veins, subtracted from the total number of teeth between these two main teeth]; DFT is the Distance from the petiole to the First Tooth on the leaf base [in millimetres]; LTW is Leaf Tip Width [one quarter of the distance between the apex and the leaf base in millimetres]). We measured five leaves per individual for 944 trees.

Results and Discussion

The STRUCTURE analysis of microsatellite data showed distinct genetic clusters for the two species, identifying 780 trees as *B. pubescens* and 164 trees as *B. pendula* (Fig. 3.1). Only five individuals could not be identified with a posterior probability of over 99% (Fig. 3.1), but even for these the posterior probability of belonging to their species was over 90%. Our STRUCTURE analysis in Wang *et al.* (2014), run under an admixture model, showed some evidence for introgressive hybridisation between the two species, but the present study shows that this small amount of introgression does not prevent clear species discriminations using the same data set. While we did not find any absolutely diagnostic species marker alleles, Table 3.1 shows examples of allele size ranges present at differential frequencies in *B. pubescens* and *B. pendula*.

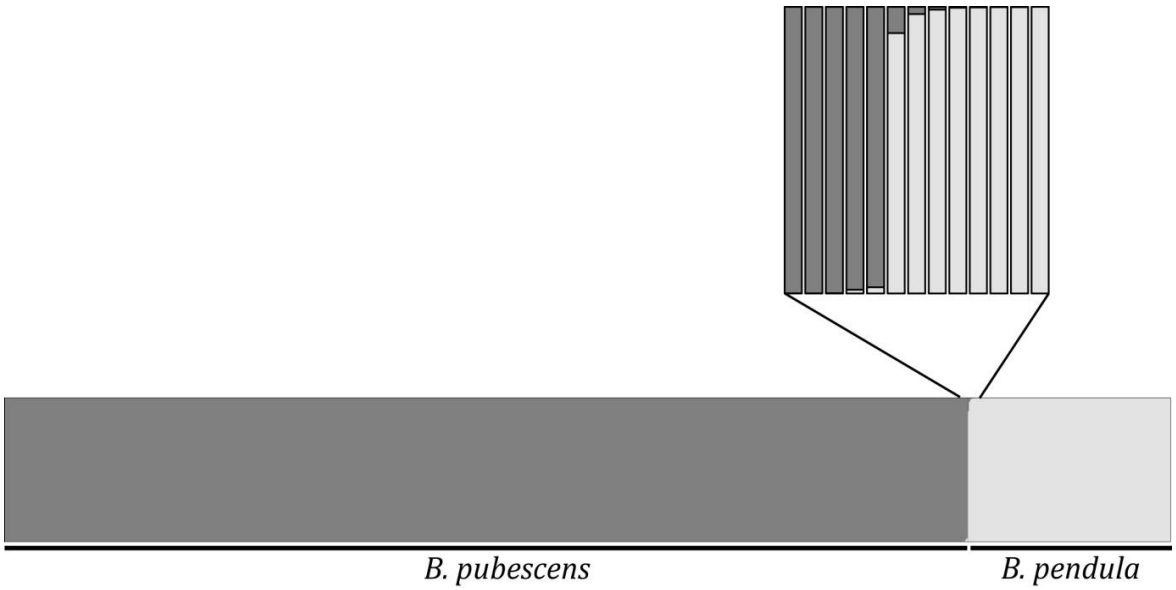


Figure 3.1 STRUCTURE analysis of 944 *Betula* trees, estimating the posterior probability that each individual is derived from each species population. Each vertical line represents an individual.

We found that trees identified by microsatellite data in STRUCTURE as *B. pubescens* had ADF scores ranging from -39.8 to 4.6, whereas *B. pendula* trees had ADF scores ranging from -17.8 to 31.4 (Fig. 3.2). No relationship was detected between the admixture values and the ADF scores (Fig. 3.3). Atkinson & Codling (1986) suggested

that an ADF score of less than zero indicates a *B. pubescens* individual, and an ADF score of more than zero indicates *B. pendula*. If we had based species identification on the ADF score alone, with zero as the boundary between *B. pendula* and *B. pubescens*, six (of 780) *B. pubescens* trees and 28 (of 164) *B. pendula* trees would be misidentified (3.6% of trees). This is a lower rate of misidentification than that found by Atkinson & Codling (1986): seven out of 56 *B. pubescens* trees were misidentified in their study on the basis of the ADF, and none of 48 *B. pendula* trees (6.7% of trees misidentified).

Table 3.1 The five microsatellite loci with the best discrimination between *B. pendula* and *B. pubescens*, showing proportion of trees from each species containing alleles within given size ranges.

Marker	Range of allele sizes (bp)	Proportion of *B. pubescens* trees containing allele(s)	Proportion of *B. pendula* trees containing allele(s)
L012	204-219	0.28	0.92
	220-246	0.27	0.04
	Both of above	0.45	0.04
L7.3	185-202	0.21	0.88
	203-238	0.30	0.06
	Alleles from both	0.49	0.06
M2.3	209-216	0.63	0.40
	217-253	0.02	0.19
	Both of above	0.35	0.41
M3.1	133-159	0.63	0.01
	160-188	0.02	0.98
	Both of above	0.35	0.01
M5.4	137-157	0.40	0.31
	158-192	0.00	0.25
	Both of above	0.60	0.45

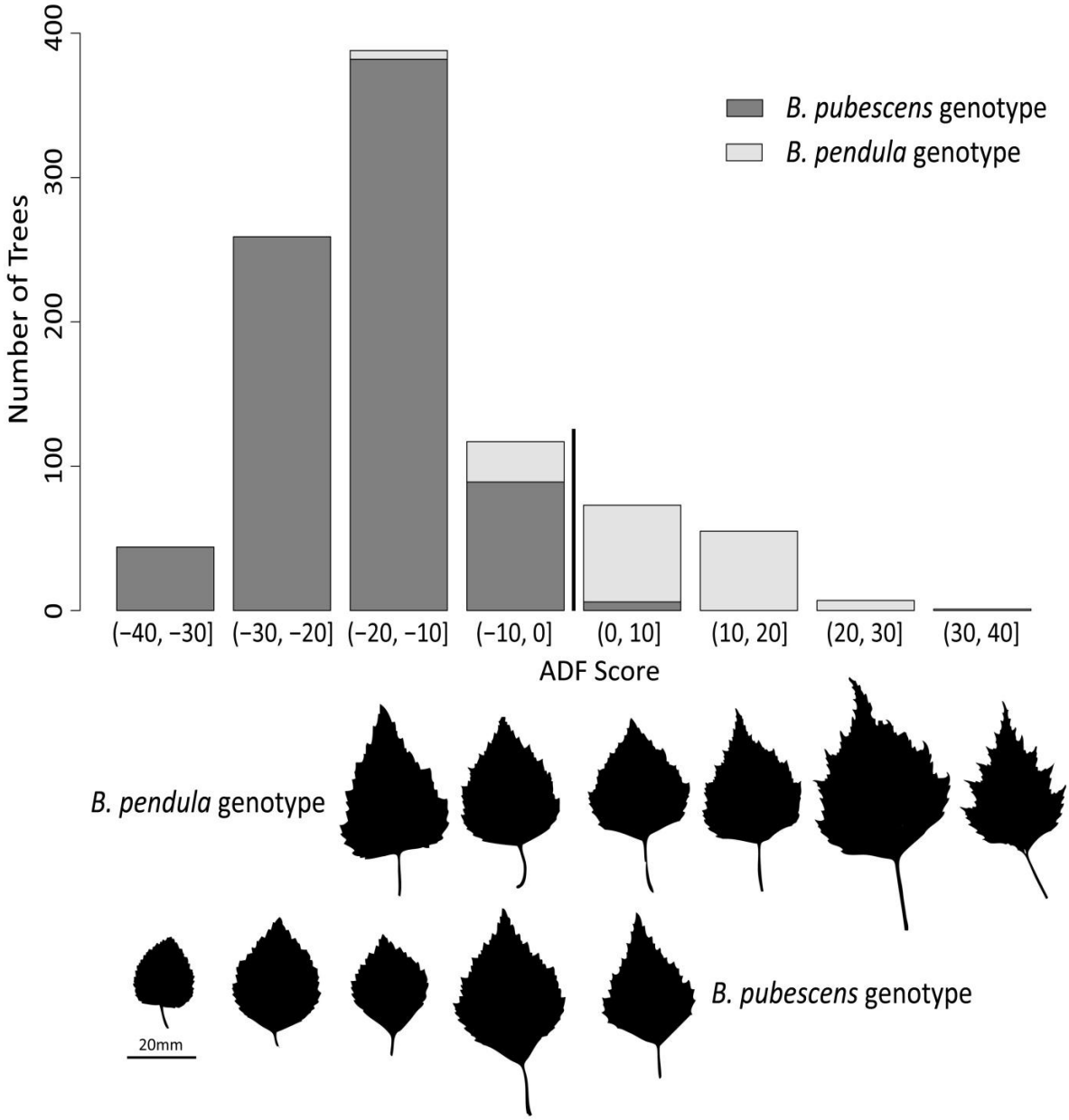


Figure 3.2 The distribution pattern of Atkinson discriminant function (ADF) scores for 944 *Betula* trees from 105 populations sampled in England, Scotland and Wales, for which microsatellite data are available. A representative leaf belonging to each ADF score range is shown below the barplot. The black line indicates the boundary between *B. pubescens* and *B. pendula* as suggested by M. D. Atkinson & A. N. Codling (1986).

We found that we could increase the success rate of our ADF species discriminations by specifying minus two, rather than zero, as the boundary between *B. pendula* and *B. pubescens*. This reduced the number of misidentifications to nine *B. pubescens* trees and 15 *B. pendula* trees, i.e. a 2.5% error rate. When the threshold was set at minus two, the majority of errors occurred in an ADF range between -11 and plus three: within this range we had an error rate of greater than 0.10. Hence, based on our results, we suggest moving the threshold to minus two and flagging -11 to plus three as a zone

of error rate over 10%. However, it should be noted that the misidentifications in the Atkinson & Codling (1986) study were all of *B. pubescens* trees with ADF scores above zero, two having ADF scores above ten. A threshold of minus two would have increased their error rate.

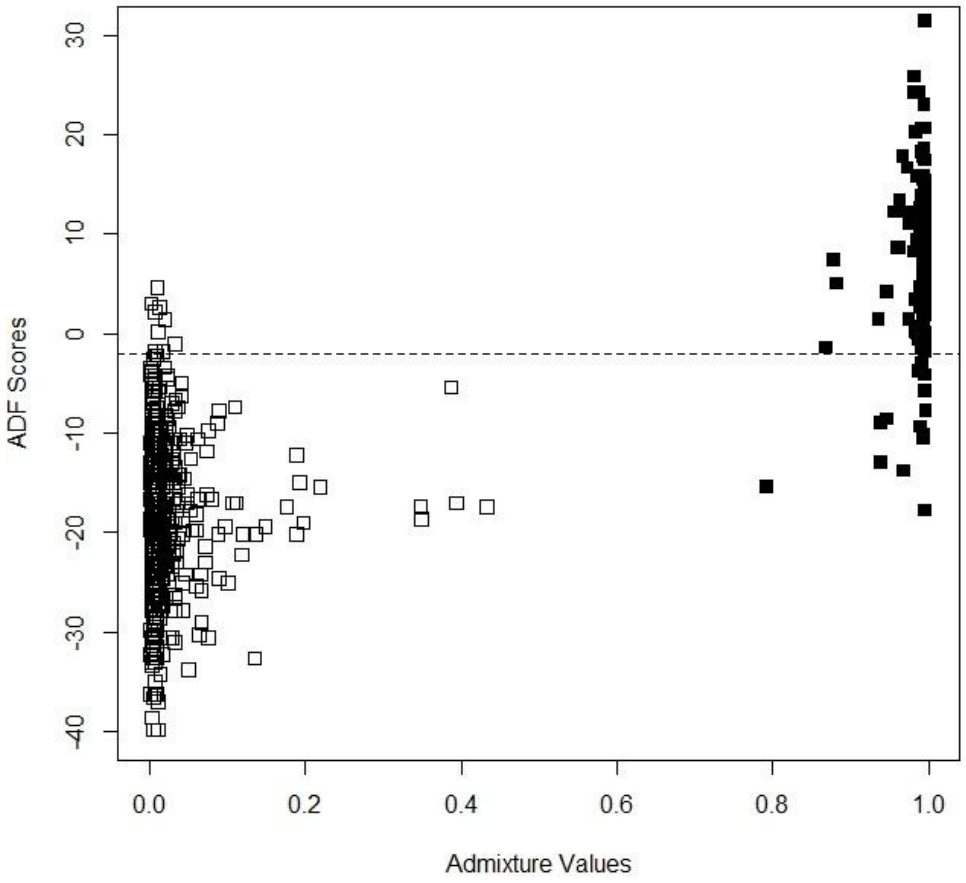


Figure 3.3 The distribution of Atkinson discriminant function (ADF) scores against the admixture values derived from STRUCTURE analysis of *B. pubescens* and *B. pendula*. The admixture value of 0 and 1 indicated 'pure' *B. pubescens* and 'pure' *B. pendula*, respectively. Open and solid squares represent *B. pubescens* individuals and *B. pendula* individuals, respectively. The dashed line at the value of minus two represented the recommended boundary line suggested between the two species.

Conclusion

The ADF is a more reliable method for discriminating between *B. pendula* and *B. pubescens* than the original paper proposing it (Atkinson & Codling, 1986) suggests. However, unlike Atkinson & Codling we found that it can misidentify both *B. pendula* and *B. pubescens* trees. We obtained maximum successful discrimination with a threshold ADF score of minus two between the two species. For population level studies the ADF score may be a useful indication of species identity, but if species identification of a particular tree is needed with certainty, microsatellite analysis, flow cytometry measurements of genome size, or chromosome counting are recommended.

Chapter 4 Molecular phylogeny and genome size evolution of the genus *Betula* (Betulaceae)

Publication information:

This chapter is based on a paper published in *Annals of Botany*, for which I was the lead author. Hugh McAllister and Paul Bartlett provided verified samples and gave instructive comments on the taxonomy of *Betula*. All authors on this manuscript contributed to editing and commenting on the original manuscript.

Wang N, McAllister HA, Bartlett PR, Buggs RJA*. 2016. Molecular phylogeny and genome size evolution of the genus *Betula* (Betulaceae). *Annals of Botany* Doi: 10.1093/aob/mcw048.

Summary

Betula L. (birch) is a genus of ~60 species, subspecies or varieties with a wide distribution in the northern hemisphere, of ecological and economic importance. A new classification of *Betula* by Ashburner and McAllister has been proposed in a recent monograph, based on morphological characters. This classification differs somewhat from previously published molecular phylogenies, which may be due to factors such as: convergent evolution, hybridisation, incomplete taxon sampling, or misidentification of samples. Whilst chromosome counts have been made for many species, few have had genome size measured. Here, we aim to produce a new phylogenetic and genome size analysis of the genus. Internal transcribed spacer (ITS) regions of nuclear ribosomal DNA were sequenced for 76 *Betula* samples verified by taxonomic experts, representing ~60 taxa of which ~24 taxa have not been included in previous phylogenetic analyses. We also sequenced a further 49 samples from other collections, and downloaded 108 ITS sequences from GenBank. Phylogenetic trees were built using these sequences. The genome sizes of 103 accessions representing nearly all described species were estimated using flow cytometry.

As expected for a gene tree of a genus where hybridisation and allopolyploidy 25 occur, our ITS tree shows clustering, but not resolved monophyly, for the morphological 26 subgenera proposed by Ashburner and McAllister. Most sections show some clustering, but species of the dwarf section *Apterocaryon* are unusually scattered. *Betula corylifolia* (subgenus *Nipponobetula*) unexpectedly clusters with species of subgenus *Aspera*. Unexpected placements are also found for *B. maximowicziana*, *B. bomiensis*, *B. nigra* and *B. grossa*. We found biogeographical disjunctions within *Betula* between Europe and N. America, and disjunctions between NE and SW Asia. The 2C-values for *Betula* ranged from 0.88 pg to 5.33 pg, and polyploids are scattered widely throughout the ITS phylogeny. Species with large genomes tended to have narrow ranges. *Betula grossa* may have formed via allopolyploidisation between parents in subgenus *Betula* and subgenus Aspera. *Betula bomiensis* may also be a wide allopolyploid. *Betula corylifolia* may be a parental species of allopolyploids in the subsection *Chinenses*. Placements of *B. maximowicziana*, *B. michauxii* and *B. nigra* need further investigation. This analysis, in line with previous studies, suggests that section *Apterocaryon* is not monophyletic and thus dwarfism has evolved repeatedly in

different lineages of *Betula*. Polyploidisation has occurred many times independently in the evolution of *Betula*.

Introduction

Phylogenetic trees based on individual genes (gene trees) provide useful data for systematics, even though the evolutionary history of a particular gene is not necessarily the same as the history of other parts of the genome, or the species (Nichols, 2001). When gene trees contradict classifications based on morphological characters two broad categories of factors can underlie this discordance. First, a gene tree may be discordant with the species tree due to the effects of hybridisation, gene duplication, polyploidy and incomplete lineage sorting (Tate & Simpson, 2003; Koonin, 2005; Degnan & Rosenberg, 2009). Second, morphological similarities may give a misleading phylogenetic signal due to convergence (Day *et al.*, 2014). In addition, specimens may be occasionally misidentified (Wiens, 2004), and insufficient sampling can be a problem when interpreting phylogenetic relationships (Pick *et al.*, 2010). Phylogenetic analysis of *Betula* L. (Betulaceae) is likely to be subject to these problems as *Betula* species are reported to frequently hybridise, include a number of polyploids and encompass several species that are similar morphologically (Ashburner & McAllister, 2013).

Betula, a genus of trees and shrubs, occupies a broad latitudinal range in the northern hemisphere from the subtropics to the arctic, populating various habitats, including bogs, highlands, tundra and forests. Species of this genus occur in natural landscapes and play important roles in horticulture and forestry (Ashburner & McAllister, 2013). Although several *Betula* species have wide ranges, some have narrow ranges and are evaluated as endangered in the IUCN Red List (Ashburner & McAllister, 2013; Shaw *et al.*, 2014). The estimated species number within the genus ranges from 30 to 120 (Furlow, 1990; Koropachinskii, 2013) and new species have been described recently (Zeng *et al.*, 2008; McAllister & Rushforth, 2011; Zeng *et al.*, 2014). The taxonomy of this genus is extremely difficult and controversial and several classifications have been proposed (Regel, 1865; Winkler, 1904; De Jong, 1993; Skvortsov, 2002). Regel (1865) divided it into subgenus *Alnaster* and subgenus *Eubetula*, with the former having the single section *Acuminatae* whereas the latter consisting of six sections (*Albae*, *Costatae*, *Dahuricae*, *Fruticosae*, *Lentae* and *Nanae*). Winkler (1904) lowered the two subgenera proposed by Regel (1865) to two sections and merged section *Dahuricae* and section *Fruticosae* of Regel (1865) into subsection *Albae*, and section *Lentae* into subsection *Costatae*. De Jong (1993) divided the genus into five subgenera: *Betula*,

Betulaster, *Betulenta*, *Chamaebetula* and *Neurobetula* (Table 4.1). Based on previous publications and specimens collected from northern Asia, Skvortsov (2002) proposed a classification of four subgenera and eight sections, namely *Sinobetula*, *Nipponobetula*, *Asperae* (sections *Asperae*, *Chinenses* and *Lentae*) and *Betula* (sections *Acuminatae*, *Betula*, *Costatae*, *Dahuricae* and *Apterocaryon*). More recently, in a monograph of *Betula* (Ashburner & McAllister, 2013), a classification into four subgenera and eight sections was proposed (Table 4.1). These subgenera are: *Nipponobetula* (section *Nipponobetula*), *Aspera* (sections *Asperae* and *Lentae*), *Acuminata* (section *Acuminatae*) and *Betula* (sections *Betula*, *Costatae*, *Dahuricae* and *Apterocaryon*) with section *Asperae* being further divided into two subsections: subsection *Asperae* and subsection *Chinenses*. This classification largely agrees with the one proposed by Skvortsov (2002), but places section *Acuminatae* (subgenus *Betula*) of Skvortsov (2002) as subgenus *Acuminata* and treats sections *Asperae*, *Chinenses* and *Lentae* of Skvortsov (2002) as subsections *Asperae*, *Chinenses* and section *Lentae*, respectively. Subgenus *Sinobetula* is not included in this recent classification since this was proposed based only on a single specimen (Skvortsov, 2002). The Ashburner and McAllister classification has not yet been evaluated phylogenetically.

Several molecular phylogenies have been published for the family Betulaceae (Bousquet *et al.*, 1992; Chen *et al.*, 1999; Forest *et al.*, 2005; Grimm & Renner, 2013) and for its constituent genera: *Alnus* (Navarro *et al.*, 2003), *Corylus* (Erdogan & Mehlenbacher, 2000; Forest & Bruneau, 2000; Whitcher & Wen, 2001), *Carpinus* (Yoo & Wen, 2002) and *Betula* (see references above). It is generally agreed that genus *Betula* is sister to *Alnus* and the remaining four genera (*Corylus*, *Carpinus*, *Ostryopsis* and *Ostrya*) form another group (Bousquet *et al.*, 1992; Chen *et al.*, 1999). Within *Betula*, current understanding of phylogenetic relationships is based primarily on five studies with only a subset of species sampled in each study (Järvinen *et al.*, 2004; Li *et al.*, 2005; Nagamitsu *et al.*, 2006; Li *et al.*, 2007; Schenk *et al.*, 2008). To our knowledge, approximately 24 taxa were not included in any previous phylogenetic studies, including *B. ashburneri*, *B. bomiensis*, *B. hainanensis* and *B. murrayana*. Some species placements in these phylogenies remain debated, such as the placement of *B. schmidtii* (Järvinen *et al.*, 2004; Li *et al.*, 2005), the grouping of *B. costata* and *B. alleghaniensis* and the placement of *B. glandulosa* within section *Asperae* (Li *et al.*, 2005).

Previous comparisons of morphological and molecular classifications in *Betula* reveal that they are partially inconsistent and contradictory (Järvinen *et al.*, 2004; Li *et al.*, 2005; Schenk *et al.*, 2008). One potential cause of this, hybridisation, is known to occur frequently between *Betula* species and has been detected based on morphological characters, molecular markers, cytogenetics and genome size analysis (Dehond & Campbell, 1989; Anamthawat-Jónsson & Tómasson, 1990; Anamthawat-Jónsson & Thórsson, 2003; Palmé *et al.*, 2004; Nagamitsu *et al.*, 2006; Karlsdottir *et al.*, 2009; Anamthawat-Jónsson *et al.*, 2010; Wang *et al.*, 2014a). It has been shown that hybridisation can occur across sections and even subgenera within *Betula* (Johnsson, 1945; Dancik & Barnes, 1972; Czernicka *et al.*, 2014; Thomson *et al.*, 2015), potentially causing discordance in phylogenetic relationships.

The recent monograph of *Betula* (Ashburner & McAllister, 2013) includes determinations of the ploidy level of *Betula* species based on chromosome counts, with levels ranging from diploid to dodecaploid and counted chromosome numbers from 2n = 28 to 2n = 168. Ploidy level is an important factor in distinguishing some of the morphologically similar species in the genus, such as diploid *B. pendula* (2n = 2x = 28) and tetraploid *B. pubescens* (2n = 4x = 56); and diploid *B. ashburneri* (2n = 2x = 28) and tetraploid *B. utilis* (2n = 4x = 56). Although ploidy level has been estimated for nearly all species of *Betula*, there are only five counts of genome size in the Plant DNA C-values Database (Bennett & Leitch, 2010), representing two diploid species, two tetraploid species and one triploid hybrid. Three of these five counts are from Anamthawat-Jónsson *et al.* (2010) where the genome size of 12 plants was measured. The genome size of another three species has been reported recently elsewhere (Bai *et al.*, 2012). Of these genome size measurements of which we are aware for *Betula*, species considered to be diploid appear to have very different genome sizes: the 2C-values of diploid species *B. populifolia*, *B. nana* and *B. nigra* were estimated to be 0.40 pg, 0.91 pg and 2.90 pg, respectively. Hence, there is a need for complete genome size information for the genus carried out under standard conditions with reliable identification of the specimens used.

Here, we constructed a genus-level phylogeny based on the nuclear ribosomal internal transcribed spacer (ITS) region for the genus *Betula* using samples that have been verified by the authors of the recent monograph of the genus, Ashburner and McAllister, except in the case of four species where samples were obtained from three researchers highly familiar with them. We used ITS because its high level of

polymorphism can help to distinguish species for phylogenetic analyses (Álvarez & Wendel, 2003) although it may suffer from complicating factors such as pseudogenes and biparental signals in recent hybrids (Razafimandimbison *et al.*, 2004). We also conducted broader analyses with samples from living collections or GenBank that have not been previously verified by the monographers, some of which were included in previous phylogenetic studies. We also measured the genome size of each taxon using flow cytometry.

Table 4.1 Various classification systems of *Betula*.

Species	Regel (1865)		Winkler (1904)		de Jong (1993)	Skvortsov (2002)		Ashburner&McAllister (2013)	
	Subgenus	Section	Section	Subsection	Subgenus	Subgenus	Section	Subgenus	Section
B. alnoides	*Alnaster*	*Acuminatae*	*Betulaster*	*Acuminatae*	*Betulaster*	*Betula*	*Acuminatae*	*Acuminata*	*Acuminatae*
B. cylindrostachya	*Alnaster*	*Acuminatae*	*Betulaster*	*Acuminatae*	*Betulaster*	*Betula*	*Acuminatae*	*Acuminata*	*Acuminatae*
B. hainanensis	—	—	—	—	—	—	—	*Acuminata*	*Acuminata*
B. luminifera	—	—	*Betulaster*	*Acuminatae*	*Betulaster*	*Betula*	*Acuminatae*	*Acuminata*	*Acuminatae*
B. maximowicziana	*Alnaster*	*Acuminatae*	*Betulaster*	*Acuminatae*	*Betulaster*	*Betula*	*Acuminatae*	*Acuminata*	*Acuminatae*
B. bomiensis	—	—	—	—	—	—	—	*Aspera*	*Asperae*
B. calcicola	—	—	—	—	*Neurobetula*	*Asperae*	*Asperae*	*Aspera*	*Asperae*
B. chichibuensis	—	—	—	—	*Neurobetula*	*Asperae*	*Asperae*	*Aspera*	*Asperae*
B. delavayi	—	—	*Eubetula*	*Costatae*	*Neurobetula*	*Asperae*	*Chinenses*	*Aspera*	*Asperae*
B. potaninii	—	—	—	—	*Neurobetula*	*Asperae*	*Asperae*	*Aspera*	*Asperae*
B. schmidtii	*Eubetula*	*Costatae*	*Eubetula*	*Costatae*	*Neurobetula*	*Asperae*	*Asperae*	*Aspera*	*Asperae*
B. chinensis	—	—	—	—	—	*Asperae*	*Chinenses*	*Aspera*	*Asperae*
B. fargesii	—	—	—	—	—	*Asperae*	*Asperae*	*Aspera*	*Asperae*
B. globispica	—	—	*Eubetula*	*Costatae*	*Betulenta*	*Asperae*	*Chinenses*	*Aspera*	*Asperae*
B. alleghaniensis	—	—	*Eubetula*	*Costatae*	*Betulenta*	*Asperae*	*Lentae*	*Aspera*	*Lentae*
B. grossa	*Eubetula*	*Costatae*	*Eubetula*	*Costatae*	*Neurobetula*	*Asperae*	*Lentae*	*Aspera*	*Lentae*
B. lenta	*Eubetula*	*Lentae*	*Eubetula*	*Costatae*	*Betulenta*	*Asperae*	*Lentae*	*Aspera*	*Lentae*
B. lenta f. uber	—	—	—	—	*Betulenta*	—	—	*Aspera*	*Lentae*
B. medwediewii	—	—	*Eubetula*	*Costatae*	*Betulenta*	*Asperae*	*Lentae*	*Aspera*	*Lentae*
B. megrelica	—	—	*Eubetula*	*Costatae*	*Betulenta*	—	—	*Aspera*	*Lentae*
B. murrayana	—	—	—	—	—	—	—	*Aspera*	*Lentae*
B. insignis ssp.	—	—	*Eubetula*	*Costatae*	*Betulenta*	*Asperae*	*Lentae*	*Aspera*	*Lentae*
B. costata	*Eubetula*	*Costatae*	*Eubetula*	*Costatae*	*Neurobetula*	*Betula*	*Costatae*	*Betula*	*Costatae*
B. ermanii	*Eubetula*	*Costatae*	*Eubetula*	*Costatae*	*Neurobetula*	*Betula*	*Costatae*	*Betula*	*Costatae*

B. ermanii var. *lanata*	—	—	*Eubetula*	*Costatae*	*Neurobetula*	—	—	*Betula*	*Costatae*
B. ashburneri	—	—	—	—	—	—	—	*Betula*	*Costatae*
B. albosinensis	—	—	—	—	*Neurobetula*	*Betula*	*Costatae*	*Betula*	*Costatae*
B. utilis	—	—	*Eubetula*	*Costatae*	*Neurobetula*	*Betula*	*Costatae*	*Betula*	*Costatae*
B. apoiensis	—	—	—	—	—	—	—	*Betula*	*Apterocaryon*
B. fruticosa	*Eubetula*	*Fruticosae*	*Eubetula*	*Albae*	*Chamaebetula*	*Betula*	*Apterocaryon*	*Betula*	*Apterocaryon*
B. glandulosa	*Eubetula*	*Nanae*	*Eubetula*	*Nanae*	*Chamaebetula*	*Betula*	*Apterocaryon*	*Betula*	*Apterocaryon*
B. humilis	*Eubetula*	*Nanae*	*Eubetula*	*Nanae*	*Chamaebetula*	—	—	*Betula*	*Apterocaryon*
B. michauxii	*Eubetula*	*Nanae*	*Eubetula*	*Nanae*	—	*Betula*	*Apterocaryon*	*Betula*	*Apterocaryon*
B. nana	*Eubetula*	*Nanae*	*Eubetula*	*Nanae*	*Chamaebetula*	*Betula*	*Apterocaryon*	*Betula*	*Apterocaryon*
B. ovalifolia	—	—	—	—	—	—	*Apterocaryon*	*Betula*	*Apterocaryon*
B. pumila	*Eubetula*	*Nanae*	*Eubetula*	*Nanae*	*Chamaebetula*	*Betula*	*Apterocaryon*	*Betula*	*Apterocaryon*
B. dahurica	*Eubetula*	*Dahuricae*	*Eubetula*	*Costatae*	*Neurobetula*	*Betula*	*Dahuricae*	*Betula*	*Dahuricae*
B. nigra	*Eubetula*	*Costatae*	*Eubetula*	*Costatae*	*Neurobetula*	*Betula*	*Dahuricae*	*Betula*	*Dahuricae*
B. raddeana	—	—	*Eubetula*	*Costatae*	*Neurobetula*	*Betula*	*Betula*	*Betula*	*Dahuricae*
B. cordifolia	*Eubetula*	*Albae*	*Eubetula*	*Albae*	*Betula*	*Betula*	*Betula*	*Betula*	*Betula*
B. halophila	—	—	—	—	—	—	—	*Betula*	*Betula*
B. occidentalis	*Eubetula*	*Albae*	*Eubetula*	*Albae*	*Betula*	*Betula*	*Betula*	*Betula*	*Betula*
B. papyrifera	*Eubetula*	*Albae*	*Eubetula*	*Albae*	*Betula*	*Betula*	*Betula*	*Betula*	*Betula*
B. pendula	*Eubetula*	*Albae*	*Eubetula*	*Albae*	*Betula*	*Betula*	*Betula*	*Betula*	*Betula*
B. populifolia	*Eubetula*	*Albae*	*Eubetula*	*Albae*	*Betula*	*Betula*	*Betula*	*Betula*	*Betula*
B. pubescens	*Eubetula*	*Albae*	*Eubetula*	*Albae*	*Betula*	*Betula*	*Betula*	*Betula*	*Betula*
B. tianshanica	—	—	—	—	—	*Betula*	*Betula*	*Betula*	*Betula*
B. corylifolia	*Eubetula*	*Costatae*	*Eubetula*	*Costatae*	*Betulenta*	*Nipponobetula*	—	*Nipponobetula*	*Nipponobetula*

Materials and Methods

Taxon sampling

In order to ensure a complete correspondence between Ashburner and McAllister's (2013) species names and the taxa included in this study, we obtained species from the Stone Lane Gardens in Devon (SL hereafter) and University of Liverpool Botanic Gardens at Ness (N hereafter) since these have been collected and curated by Ashburner and McAllister. In addition, we obtained four species (*B. alnoides*, *B. delavayi*, *B. glandulosa* and *B. hainanensis*) from Jie Zeng (Institute of Tropical Forestry, Chinese Academy of Forestry), Paul Grogan (Queen's University, Canada) and Zhikun Wu (Kunming Institute of Botany, Chinese Academy of Sciences) who have studied them over many years. We built our main phylogenetic tree using these verified species. We then also built a phylogenetic tree including additional samples obtained from the Royal Botanic Gardens Kew, the Royal Botanic Garden Edinburgh, the Helsinki Botanic Garden (Table 4.2), field collections and GenBank sequences from previous published phylogenetic analyses.

DNA extraction, amplification and sequencing

Genomic DNA was isolated from silica-dried cambial tissue (green vascular tissue located beneath the outer bark of woody stems) or leaves following a modified 2× CTAB (cetyltrimethylammonium bromide) protocol (Wang *et al.*, 2013). The isolated DNA was assessed with 1.0% agarose gels and measured with Qubit 2.0 Fluorometer (Invitrogen, Life technologies) using Broad-range assay reagents. The quantified DNA was then diluted to a final concentration of 10-20 ng/μl for subsequent use. The nuclear ribosomal internal transcribed spacer (nrITS) region (ITS1, 5.8s and ITS2) was amplified using primers ITS4 (White *et al.*, 1990) and ITSLeu (Baum *et al.*, 1998). The volume of reaction mix was 20 μl containing: 0.4 μl AmpliTaq polymerase, 2.0 μl 10 × NH_4 buffer (Bioline), 1.6 μl 50 mM $MgCl_2$ (Bioline), 0.5 μl 100 mM dNTP, 0.8 μl of each primer (10 mM), 12.9 μl ddH_2O and 1 μl diluted DNA (10-20 ng). The PCR was carried out using a touchdown program, consisting of an initial denaturation at 95°C for 3 min, followed by 32 cycles of 1 min at 94°C, 50 s at 56-52°C, 1.5 min at 72°C, and ended with an extension step of 10 min at 72°C. The PCR products were purified by binding a 0.8 volume of Ampure beads (Beckman Coulter Inc.). The purified PCR products were diluted to ~20 ng/μl in ddH_2O prior to sending them to

Eurofins (Ebersberg, Germany) for sequencing.

Phylogenetic analyses

ITS tree based on verified samples — Seventy-six fully verified accessions representing ~60 *Betula* species and various subspecies, varieties and natural hybrids were sequenced. The ITS sequences of these verified accessions were checked for recombination with the RDP4 program (Martin *et al.*, 2015) using seven automated detection methods: Bootscanning (Salminen *et al.*, 1995); Chimaera (Posada & Crandall, 2001); GENECONV (Padidam *et al.*, 1999); MaxChi (Smith, 1992); RDP (Martin *et al.*, 2005); SiScan (Gibbs *et al.*, 2000); and 3SEQ (Boni *et al.*, 2007). No signal of recombination was detected using all these methods. We downloaded ITS sequences of nine species from other genera of Betulaceae from GenBank, for use as outgroup taxa. In total, 85 sequences were aligned using BioEdit v 7.0.9.0 with default parameters and the alignment was edited manually where necessary. A maximum likelihood (ML) analysis was conducted in PhyML v. 3.0 with the default settings (Guindon and Gascuel 2003) and with the best-fit substitution model GTR + G, as selected in jModelTest2.0 (Guindon & Gascuel, 2003; Darriba *et al.*, 2012) using the Akaike Information Criterion (AIC). A Bayesian inference (BI) analysis was also conducted using the program MrBayes v.3.2 (Ronquist *et al.*, 2012). Two independent runs were performed. For each run, ten million generations were completed with four chains (three heated, one cold). Trees were sampled every 1000th generation and the first 25% of runs were discarded as burn-in. Convergence was assessed by determining that the average standard deviation of split frequencies reached a value of below 0.01. A majority-rule consensus of the remaining trees from the two runs was produced and used as the Bayesian inference tree with posterior probabilities (PP).

ITS tree based on all samples — In addition to the verified accessions, another 49 accessions were sequenced (Table 4.2) and ninety-nine sequences of *Betula* species were retrieved from GenBank. A total of 233 sequences were aligned and analyzed with ML and BI as described above. The consensus trees generated using the above methods were visualised in FigTree v.1.3.1 (http://tree.bio.ed.ac.uk/software/figtree) and edited in Adobe Illustrator CS4 (Adobe Systems).

ITS tree based on diploid samples — We also conducted phylogenetic analyses exclusively on verified species that our C-value measurements (see below) showed to be diploid. Thirty-three *Betula* accessions were included. A ML tree was conducted using the same parameters as described above.

Genome size analysis

We measured the genome size of nearly all samples collected from SL and N to correlate them with ploidy levels obtained from chromosome counts. Fresh leaves or cambial tissue were co-chopped with internal standards: *Oryza sativa* 'IR36' (Bennett & Smith, 1991), *Solanum lycopersicum* L. 'Stupiké polní rané' (Doležel *et al.*, 1998), *Petroselinum crispum* (Mill.) Nyman ex A.W.Hill "Champion Moss Curled" (Obermayer *et al.*, 2002) and *Pisum sativum* L. "Minerva Maple" (Bennett & Smith, 1991) in 1 ml Extraction Buffer (Cystain PI absolute P, Partec GMBH) and then filtered into the tube containing 2.0 ml Staining Solution (Cystain PI absolute P, Partec GMBH) with 12 µl propidium iodide (PI). These samples were incubated at room temperature in the dark for about 30 mins. Three to five replicates per sample with each replicate including over 5000 nuclei were analyzed in a Partec CyFlow Space flow cytometer (Partec, GmbH, Germany) fitted with a 100-mW green solid state laser (Cobolt Samba; Cobolt, Sweden). Four taxa were analyzed with less than three replicates (Table 4.2). The resulting histograms were analyzed with the Flow-Max software (v. 2.4, Partec GmbH).

The ranges of the species for which we measured genome size were divided into four loose categories: narrow (species occurring in a single or a few localities and tending to be endangered), medium (species occurring commonly in multiple areas), widespread (species occupying several parts of a continent) and very widespread (species spread extensively within a continent or across continents) (Table 4.3) based on distribution information in the recent monograph of *Betula* (Ashburner & McAllister, 2013). For species in which multiple individuals were measured, the mean genome size was used for subsequent analysis. Using the average ploidy level and the mean 2C-value of each range category, statistically significant differences between categories were tested using analysis of variance (ANOVA). Tukey HSD *post-hoc* test was performed at $P < 0.05$ when results of ANOVA indicated significance ($\alpha \leq 0.05$). All analysis and plots were performed in R 3.1.0 (R Develoment Core Team, 2012) and the package 'ggplot2' (Wickham, 2009).

Table 4.2 Detailed information of the taxa used for ITS sequencing and taxa used for genome size estimation.

Species (Ploidy level) and ITS sequence GenBank accession number[1]	Genome Size (s.d.[2])/pg	Living collection[3]	Native range	Herbarium accession numbers[4]
B. alnoides Buchanan-Hamilton ex D. Don (4x) KT308940	1.95 (0.01)	n/a	Guangxi, China	BM001122936
B. cylindrostachya Lindl. ex Wall (4x) KT308941	1.91 (0.01)	SL	India	BM001122961
B. hainanensis J. Zeng, B.Q. Ren, J.Y. Zhu & Z.D. Chen (2x) KT308942	0.91 (0)	n/a	Hainan, China	BM001122937
B. luminifera H.Winkl. (2x) KT308944	1.00 (0.01)	RBGE	Yunnan, China	E 19933472 G
B. luminifera H.Winkl. KT308943		N	Sichuan, China	BM001123047
B. luminifera H.Winkl. KT308939		K	Sichuan, China	
B. maximowicziana Regel (2x) KT308945	0.93 (0)	SL	Japan	BM001122968
B. maximowicziana Regel (2x) KT308946	0.96 (0)	N	Japan	BM001123022
B. bomiensis P.C.Li (4x) KT308911	2.20 (0)	N	Tibet, China	BM001123011
B. bomiensis P.C.Li KT308912		RBGE	Tajikistan	E 20110653 A
B. calcicola (W.W.Sm.) P.C.Li (2x) KT308914	0.91 (0.01)	N	Yunnan, China	BM001123012
B. chichibuensis Hara (2x) KT308916	0.92 (0.01)	SL	Japan	BM001122959
B. chichibuensis Hara (2x) KT308915	0.91 (NA)	SL	Japan	BM001122958
B. delavayi Franch. (6x) KT308921	3.20 (0.01)	SL	Yunnan, China	BM001122963
B. delavayi Franch. KT308913		n/a	Yunnan, China	BM001122938
***B. delavayi* Franch.** KT308922		K		
B. potaninii Batalin (2x) KT308909	1.08 (0)	N	Sichuan, China	BM001123030
B. potaninii Batalin KT308910		K		
B. schmidtii Regel (2x) KT308920	0.92 (0)	N	Russian Far East	BM001123034
B. schmidtii Regel KT308919		K		
***B. skvortsovii* McAll. & Ashburner (2x)** KT308961	1.00 (0)	n/a		
B. chinensis Maxim. (6x) KT308917	2.76 (0.01)	N	S. Korea	BM001123013

B. chinensis Maxim. (8x) KT308918	3.12 (0.03)	N	S. Korea	BM001123014
B. fargesii (Franchet) P. C. Li. (10x) KT308906	5.17 (0.01)	N	Hubei, China	BM001123019
B. globispica Shirai (10x) KT308905	4.88 (0.03)	N	Japan	BM001123020
B. globispica Shirai KT308904		K		
B. alleghaniensis Britton (6x) KT308925	2.97 (0.01)	N	Tenessee, USA	BM001123048
B. grossa Siebold & Zucc. (6x) KT308934	2.58 (0)	SL	Honshu, Japan	BM001122965
B. grossa Siebold & Zucc. KT308935		K	Honshu, Japan	
B. lenta L. (2x) KT308936	0.95 (0)	SL	Vermont, USA	BM001122993
B. lenta f. *uber* (Ashe) Fernald (2x) KT308938	0.96 (0)	SL	Virginia, USA	BM001122994
B. lenta f. *uber* (Ashe) Fernald KT308937		K		
B. medwediewii Regel (10x) KT308931	4.73 (0.02)	SL	Caucasus, Georgia	BM001123004
B. medwediewii Regel KT308930		K		
B. medwediewii Regel (10x) ITS not sequenced	4.78 (0.02)	N	Caucasus, Georgia	BM001123023
B. megrelica D. Sosn. (12x) KT308933	5.12 (0.01)	SL	Caucasus, Georgia	BM001122969
B. megrelica D. Sosn. KT308932		K		
B. murrayana B. V. Barnes & Dancik (8x) KT308926	3.03 (0.01)	N	Ontario, Canada	BM001123026
B. insignis Franch. (10x) KT308927	4.71 (0.01)	SL	Yunnan, China	BM001122967
B. insignis Franch. KT308928		RBGE	Guizhou, China	E 20050415 R
B. insignis Franch. (10x) ITS not sequenced	4.44 (0.02)	n/a	Guizhou, China	BM001122940
B. insignis ssp. *fansipanensis* Ashburner & McAll. (10x) KT308929	5.33 (0.01)	n/a	Yunnan, China	BM001122941
B. costata Trautv. (2x) KT308958	0.93 (0)	N	Beijing, China	BM001123016
B. ermanii Cham. (4x) KT308956	2.00 (0.01)	SL	Hokkaido, Japan	BM001122964
B. ermanii Cham. KT308957		K		
B. ermanii var. *lanata* Regel (4x) KT30860	2.12 (0)	N	Russian Far East	BM001123018
B. ermanii var. *lanata* Regel KT308959		H	Russia	BM001122957
B. ashburneri McAllister & Rushforth (2x) KT308953	0.98 (0)	SL	SE Tibet, China	BM001122997

B. ashburneri McAllister & Rushforth KT308952		RBGE	Bhutan	E 19841878 A
B. ashburneri McAllister & Rushforth (2x) ITS not sequenced	0.99 (0)	SL	Shanxi, China	BM001122998
B. ashburneri McAllister & Rushforth (2x) ITS not sequenced	0.98 (0.01)	SL	Nepal	BM001123006
***B. albosinensis* Burkill (4x)** KT308924		H		BM001122956
B. albosinensis Burkill (4x) KT308954	2.06 (0.04)	N	China	BM001123036
B. albosinensis Burkill var. *septentrionalis* C. K. Schneider (4x) KT308947	2.04 (0)	SL	Sichuan, China	BM001122996
B. utilis D.Don (4x) KT308948	2.12 (0)	N	Nepal	BM001123035
B. utilis D.Don (4x) KT308949	2.15 (0)	SL	Sichuan, China	BM001123005
***B. utilis* D.Don var. *occidentalis* Ashburner & A.D.Schill. (4x)** KT308923	1.78 (NA)	RBGE	Tajikistan	E 20110849 A
B. utilis D.Don var. *occidentalis* Ashburner & A.D.Schill. KT308950		K		
B. utilis D.Don var. *prattii* Burkill (4x) KT308955		K		
B. utilis D.Don var. *prattii* Burkill (4x) ITS not sequenced	2.10 (0.01)	N	SW China	BM001123044
B. utilis D. Don var. *jacquemontii* (Spach) Winkle (4x) KT308951	2.09 (0)	SL	Nepal	BM001122995
B. utilis D. Don var. *jacquemontii* (Spach) Winkle 'Doorenbos' (4x) ITS not sequenced	1.89 (0.01)	SL	cultivar	
B. utilis D. Don var. *jacquemontii* (Spach) Winkle 'Graywood Ghost' (4x) ITS not sequenced	2.16 (0.01)	SL	cultivar	
B. utilis D. Don var. *jacquemontii* (Spach) Winkle 'Gregory' (4x) ITS not sequenced	2.13 (0.02)	SL	cultivar	
B. utilis D. Don var. *jacquemontii* (Spach) Winkle 'Hergest' (4x) ITS not sequenced	1.98 (0)	SL	cultivar	
B. utilis D. Don var. *jacquemontii* (Spach) Winkle 'Knight' (4x) ITS not sequenced	2.07 (0.01)	SL	cultivar	
B. utilis D. Don var. *jacquemontii* (Spach) Winkle 'Long Trunk' (4x) ITS not sequenced	1.98 (0.01)	SL	cultivar	
B. utilis D. Don var. *jacquemontii* (Spach) Winkle 'Sauwola white' (4x) ITS not sequenced	1.93 (0.02)	SL	cultivar	
***B. turkstanica* Litv.** KT308992		K	Tajikistan	

B. apoiensis Nakai ex H.Hara (4x) ITS not sequenced	2.07 (0.01)	SL		BM001123007
***B. fruticosa* Pall.** KT309002		H	Russia	BM001122955
B. glandulosa Michaux KT309017		n/a	Canada	BM001122942
***B. glandulosa* Michaux** KT308995		K	Irkutsk, Russia	
B. humilis Schrank (2x) KT309026	0.98 (0.01)	N	Romania	BM001123021
B. humilis Schrank (2x) KT309024	0.94 (0.01)	SL	Poland	BM001122966
B. humilis Schrank KT309025		K		
B. michauxii Spach (2x) KT308978	0.95 (0)	N	Canada	BM001123024
B. middendorffii Trautv. & C.A.Mey (4x) KT308986	2.06 (NA)	n/a	Russian Far East	BM001123049
B. nana L. (2x) KT309018	1.00 (0)	n/a	Scotland	BM001122943
B. nana L. (2x) KT309020	0.92 (0.01)	n/a	Scotland	BM001074532
***B. nana* ssp. *exilis* (Sukaczev) Hultén** KT309019		H	Canada	BM001122954
B. ovalifolia Ruprecht (4x) KT309022	1.92 (0.01)	SL	Mongolia	BM001122972
B. ovalifolia Ruprecht KT309023		K		
B. pumila L. (4x) KT309021	2.10 (0.01)	SL	Canada	BM001122991
B. dahurica Pall. (6x) KT308962	3.60 (0.02)	SL	Hokkaido, Japan	BM001122962
B. dahurica Pall. (6x) ITS not sequenced	3.79 (0.02)	N	Russian Far East	BM001123039
B. dahurica Pall. (8x) KT308963	4.57 (0)	N	Hokkaido, Japan	BM001123017
B. dahurica Pall. (8x) ITS not sequenced	4.36 (0.04)	N	S. Korea	BM001123040
B. dahurica Pall. (8x) ITS not sequenced	4.48 (0.02)	N	Russian Far East	BM001123041
B. dahurica Pall. (8x) ITS not sequenced	4.53 (0.02)	N	Nobeyama, Japan	BM001123042
B. dahurica Pall. (8x) ITS not sequenced	4.45 (0.01)	N	Russian Far East	BM001123043
B. nigra L. (2x) KT308964	0.88 (0)	SL	USA	BM001122970
B. nigra L. KT308965		K		

B. raddeana Trautv. (6x) KT308966	2.84 (0)	SL	Georgia	BM001122992
***B. browicziana* Güner** KT308968		RBGE	Turkey	E 20081535 C
B. cordifolia Regel (2x) KT309016	0.96 (0)	N	Canada	BM001123015
B. cordifolia Regel (2x) ITS not sequenced	1.00 (0)	SL	Canada	BM001122960
B. cordifolia Regel KT309015		RBGE	USA	E 19961304 A
B. halophila Ching KT308967		n/a	Xinjiang, China	
B. microphylla Bunge (4x) KT308984	1.81 (0)	N	Mongolia	BM001123025
B. occidentalis Hooker (2x) KT309027	0.96 (0)	SL	Montana, USA	BM001122971
B. occidentalis Hooker KT309028		H	Albert, Canada	
B. papyrifera Marshall (6x) KT309011	2.94 (0.01)	SL	Ontarioo, Canada	BM001122973
B. papyrifera Marshall (6x) ITS not sequenced	2.94 (0.02)	SL	Minnosota, USA	BM001122974
B. papyrifera Marshall KT309012		K		
***B. papyrifera* Marshall** KT309013		K		
B. papyrifera Marshall var. *commutata* Regel (6x) KT309014	2.95 (0.02)	SL	Vancouver, Canada	BM001122975
B. pendula Roth ssp. *mandshurica* (Reg.) Nakai (2x) KT308996	0.95 (0)	N	Japan	BM001123028
B. pendula Roth ssp. *mandshurica* (Reg.) Nakai (2x) ITS not sequenced	0.98 (0.01)	N	Russian Far East	BM001123050
B. pendula Roth ssp. *mandshurica* (Reg.) Nakai (2x) KT309005	0.94 (0)	N	Alberta, Canada	BM001123029
B. pendula Roth ssp. *mandshurica* (Reg.) Nakai (2x) KT309008	0.93 (0.01)	SL	Hokkaido, Japan	BM001122976
B. pendula Roth ssp. *mandshurica* (Reg.) Nakai KT308990		H	Russia	BM001122952
B. pendula Roth ssp. *mandshurica* (Reg.) Nakai KT308999		H	Milkovo, Bulgaria	BM001122953
B. pendula Roth ssp. *pendula* Roth KT309006		n/a	England	BM001122944
B. pendula Roth ssp. *pendula* Roth (2x) KT309000	0.92 (0)	SL	Sicily	BM001122977
B. pendula Roth ssp. *pendula* Roth (2x) KT308997	0.91 (0)	SL	Poland	BM001122978
B. pendula Roth ssp. *pendula* Roth KT309001		H	Finland	BM001122950

B. pendula Roth ssp. *pendula* Roth KT309007		H		BM001122951
B. pendula Roth ssp. *szechuanica* Ashburner & McAll. (2x) KT309004	0.91 (0)	SL	Sichuan, China	BM001122979
B. pendula Roth ssp. *szechuanica* Ashburner & McAll. (2x) ITS not sequenced	0.99 (0.01)	SL	Sichuan, China	BM001122980
B. pendula Roth ssp. *szechuanica* Ashburner & McAll. KT308998		K	Yunnan, China	
B. pendula Roth ssp. *szechuanica* Ashburner & McAll. KT309003		K		
B. pendula Roth ssp. *szechuanica* Ashburner & McAll. ITS not sequenced	0.93 (0)	N	Sichuan, China	BM001123027
***B. resinifera* Britton** KT308991		K	USA	
B. populifolia Marshall (2x) KT309009	0.96 (0)	SL	Vermont, USA	BM001122981
B. populifolia Marshall (2x) KT309010	0.94 (0)	SL	Vermont, USA	BM001122982
B. populifolia Marshall KT308994		K		
***B. obscura* Kotula** KT308993		K		
B. pubescens Ehrh. var. *celtiberica* Rivas Mart. (4x) KT308972	1.88 (0.01)	SL	Spain	BM001122983
B. pubescens Ehrh. var. *celtiberica* Rivas Mart. KT308977		K		
B. pubescens Ehrh. var. *fragrans* Ashburner & McAll. (4x) KT308975	1.88 (0.01)	SL	Scotland	BM001122985
B. pubescens Ehrh. var. *fragrans* Ashburner & McAll. (4x) ITS not sequenced	1.88 (0)	SL	Oslo, Norway	BM001122986
B. pubescens Ehrh. var. *fragrans* Ashburner & McAll. (4x) KT308974	1.94 (0)	N	Scotland	BM001123031
B. pubescens Ehrh. var. *litiwinowii* Ashburner & McAll. (4x) KT308971	1.79 (0.01)	SL	Caucasus, Georgia	BM001122987
B. pubescens Ehrh. var. *litiwinowii* Ashburner & McAll. KT308983	1.84 (0.01)	N	Armenia	BM001123032
B. pubescens Ehrh. var. *murithii* (Gaudin ex Regel) Gremli (4x) ITS not sequenced	1.87 (0)	SL	Switzerland	BM001123010
B. pubescens Ehrh. var. *pubescens* (4x) ITS not sequenced	1.90 (0.01)	N	Turkey	BM001123046
B. pubescens Ehrh. var. *pubescens* (4x) KT308981	1.90 (0)	SL	NE Turkey	BM001122988
B. pubescens Ehrh. var. *pubescens* (4x) KT308969		n/a	England	BM001122945
B. pubescens Ehrh. var. *pubescens* (4x) KT308970		n/a	England	BM001122946

B. pubescens Ehrh. var. *pubescens* (4x) ITS not sequenced	1.91 (0.01)	n/a	England	BM001122947
B. pubescens Ehrh. var. *pubescens* KT308982		H	Czech Republic	BM001122948
B. pubescens Ehrh. var. *pumila* (L.) Govaerts (4x) KT308976	1.91 (0.01)	SL	Trondelog, Norway	BM001122989
B. pubescens Ehrh. var. *pumila* (L.) Govaerts (4x) ITS not sequenced	1.92 (0.02)	SL	Trondelog, Norway	BM001122990
B. pubescens Ehrh. var. *pumila* (L.) Govaerts KT308980		H	Finland	BM001122949
B. pubescens Ehrh. var. *pumila* (L.) Govaerts (4x) KT308973	2.01 (0.01)	N	Norway	BM001123033
B. x caerulea Blanch. (2x) KT308988	0.97 (0)	SL	Vermont, USA	BM001122999
B. x caerulea Blanch. KT308987		K		
B. x minor (Tuckerman) Fern. (2x) KT308985	0.95 (0.01)	SL	Cananda	BM001123000
B. x utahensis Britton (4x) KT308979	1.82 (0.01)	SL	Montana, USA	BM001123001
B. tianshanica Rupr. (4x) KT308989	1.90 (NA)	RBGE	China	E 20051397 A
B. corylifolia Regel & Maxim (2x) KT308908	0.97 (0)	RBGE	Japan	E 20052047 O
B. corylifolia Regel & Maxim KT308907		RBGE	Japan	E 20052047 P

[1]Bold font indicates taxa with unexpected phylogenetic positions. [2]NA indicates less than three replicates were measured for genome size analysis; a blank means the genome size of this taxon was not estimated. [3]SL: Stone Lane Gardens; N: Ness Gardens; K: Royal Botanic Gardens, Kew; RBGE: Royal Botanic Garden Edinburgh; H: Helsinki Botanic Garden in Finland; n/a: samples are not in a living collection. [4]Accession numbers starting with BM are for the Natural History Museum, London, and accession numbers starting E are for the Royal Botanic Garden, Edinburgh.

Table 4.3 Detailed information of the taxa used for comparing the average ploidy level and the mean 2C value of genome size of different ranges.

Species	2C	1C	1Cx	Ploidy level	Range[1]	Section[2]	Subgenus
B. alnoides Buchanan-Hamilton ex D. Don	1.95	0.98	0.49	4	M	*Acuminatae*	*Acuminata*
B. cylindrostachya Lindl. ex Wall	1.91	0.96	0.48	4	M	*Acuminatae*	*Acuminata*
B. hainanensis J. Zeng, B.Q. Ren, J.Y. Zhu & Z.D. Chen	0.91	0.46	0.46	2	M	*Acuminatae*	*Acuminata*
B. luminifera H.Winkl.	1.00	0.50	0.50	2	W	*Acuminatae*	*Acuminata*
B. maximowicziana Regel	0.93	0.47	0.47	2	M	*Acuminatae*	*Acuminata*
B. bomiensis P.C.Li	2.20	1.10	0.55	4	N	*Asperae*	*Aspera*
B. calcicola (W.W.Sm.) P.C.Li	0.91	0.46	0.46	2	N	*Asperae*	*Aspera*
B. chichibuensis Hara	0.91	0.46	0.46	2	M	*Asperae*	*Aspera*
B. chinensis Maxim.	2.76	1.38	0.46	6	M	*Asperae*	*Aspera*
B. chinensis Maxim.	3.12	1.56	0.39	8	M	*Asperae*	*Aspera*
B. delavayi Franch.	3.20	1.60	0.53	6	N	*Asperae*	*Aspera*
B. fargesii (Franchet) P. C. Li.	5.17	2.59	0.52	10	N	*Asperae*	*Aspera*
B. globispica Shirai	4.88	2.44	0.49	10	N	*Asperae*	*Aspera*
B. potaninii Batalin	1.08	0.54	0.54	2	N	*Asperae*	*Aspera*
B. schmidtii Regel	0.92	0.46	0.46	2	M	*Asperae*	*Aspera*
B. alleghaniensis Britton	2.97	1.49	0.50	6	M	*Lentae*	*Aspera*
B. grossa Siebold & Zucc.	2.58	1.29	0.43	6	M	*Lentae*	*Aspera*
B. insignis Franch.	4.71	2.36	0.47	10	M	*Lentae*	*Aspera*
B. insignis ssp. *fansipanensis* Ashburner & McAll.	5.33	2.67	0.53	10	N	*Lentae*	*Aspera*
B. lenta L.	0.96	0.48	0.48	2	M	*Lentae*	*Aspera*
B. medwediewii Regel	4.73	2.37	0.47	10	N	*Lentae*	*Aspera*
B. megrelica D. Sosn.	5.12	2.56	0.43	12	N	*Lentae*	*Aspera*
B. murrayana B. V. Barnes & Dancik	3.03	1.52	0.38	8	N	*Lentae*	*Aspera*

B. humilis Schrank	0.94	0.47	0.47	2	VW	*Apterocaryon*	*Betula*
B. michauxii Spach	0.95	0.48	0.48	2	M	*Apterocaryon*	*Betula*
B. nana L.	1.00	0.50	0.50	2	VW	*Apterocaryon*	*Betula*
B. ovalifolia Ruprecht	1.92	0.96	0.48	4	M	*Apterocaryon*	*Betula*
B. pumila L.	2.10	1.05	0.53	4	W	*Apterocaryon*	*Betula*
B. cordifolia Regel	0.96	0.48	0.48	2	W	*Betula*	*Betula*
B. microphylla Bunge	1.81	0.91	0.45	4	M	*Betula*	*Betula*
B. occidentalis Hooker	0.96	0.48	0.48	2	VW	*Betula*	*Betula*
B. papyrifera Marshall	2.95	1.47	0.49	6	VW	*Betula*	*Betula*
B. pendula Roth ssp. *pendula* Roth	0.91	0.46	0.46	2	VW	*Betula*	*Betula*
B. populifolia Marshall	0.94	0.47	0.47	2	W	*Betula*	*Betula*
B. pubescens Ehrh. var. *pubescens*	1.88	0.95	0.48	4	VW	*Betula*	*Betula*
B. tianshanica Rupr.	1.90	0.95	0.48	4	M	*Betula*	*Betula*
B. ashburneri McAllister & Rushforth	0.99	0.50	0.50	2	M	*Costatae*	*Betula*
B. costata Trautv.	0.93	0.47	0.47	2	M	*Costatae*	*Betula*
B. ermanii Cham.	2.06	1.00	0.50	4	W	*Costatae*	*Betula*
B. utilis D.Don	2.08	1.02	0.51	4	VW	*Costatae*	*Betula*
B. dahurica Pall.	3.60	1.80	0.60	6	M	*Dahuricae*	*Betula*
B. dahurica Pall.	4.57	2.29	0.57	8	M	*Dahuricae*	*Betula*
B. nigra L.	0.88	0.44	0.44	2	M	*Dahuricae*	*Betula*
B. raddeana Trautv.	2.84	1.42	0.47	6	N	*Dahuricae*	*Betula*
B. corylifolia Regel & Maxim	0.97	0.49	0.49	2	N	*Nipponobetula*	*Nipponobetula*

[1]N, M, W and VW indicate narrow (species occurs in a single or a few localities and tend to be endangered), medium (species occurs commonly in multiple areas), widespread (species spreading within some parts of a continent) and very widespread (species spreading within a continent or across continents) ranges, respectively. [2]Species were classified according to Ashburner and McAllister's classification.

To further investigate the evolution of genome size in *Betula*, we calculated the monoploid genome size, 1Cx (found by dividing the 2C-value by the ploidy level of the species) (Greilhuber *et al.*, 2005), for each of the 71 verified accessions plus each accession of *B. pubescens* and *B. tianshanica* from RBGE. These 1Cx-values were grouped according to the ITS clade membership of the species; for each group, 1Cx-values were plotted against ploidy level. We also compared the homogeneity of variance for 1Cx-values among diploid (2x), tetraploid (4x), hexaploid (6x) and octoploid and above (8x-12x) accessions, with R package 'lawstat' using the modified robust Brown-Forsythe Levene-type test with 1000 bootstraps (Hui *et al.*, 2008).

Results

The phylogeny of "verified" **Betula** *species based on ITS*

The aligned ITS data matrix for "verified" sample-set contains 85 ITS sequences and 618 characters of which 157 characters are variable and 111 informative. There is broad agreement between our ML (Fig. 4.1) and Bayesian analyses; below we discuss our results based on the ML analysis as these give greater resolution. To facilitate discussion we have labeled six main clades. Clades I, II and III consist of species of section *Lentae* (subgenus *Aspera*). *Betula alleghaniensis* is sister to *B. murrayana* whereas *B. insignis* is sister to *B. insignis* ssp. *fansipanensis* forming clade I and II, respectively. Clade III consists of *B. lenta*, *B. megrelica* and *B.medwediewii*. Clade IV includes species of section *Asperae* and *B. corylifolia*, the single species of subgenus *Nipponobetula*, which appears to be sister to *Aspera* subsection *chinensis*. Clade V contains all species of the subgenus *Acuminata* together with a sub-clade of *B. bomiensis* (subsection *Asperae*) and *B. nigra* (section *Dahuricae*), the latter being on a long branch. Clade VI contains all but one of the species in subgenus *Betula* plus *B. grossa* (subgenus *Aspera*, section *Lentae*) and *B. maximowicziana* (subgenus *Acuminata*). The only species of subgenus *Betula* not found in Clade VI is *B. michauxii*, which forms a polytomy with clades IV, V and VI. Within Clade VI, the various sections of subgenus *Betula* do not form unique sub-clades, athough *B. costata*, *B. utilis* and *B. ashburneri* from section *Costatae* cluster together and *B. pubescens*, *B. pendula* and their subspecies/varieties cluster together (Fig. 4.1). Phylogenetic relationships within the above clades are not fully resolved.

The phylogeny of all available **Betula** *ITS sequences*

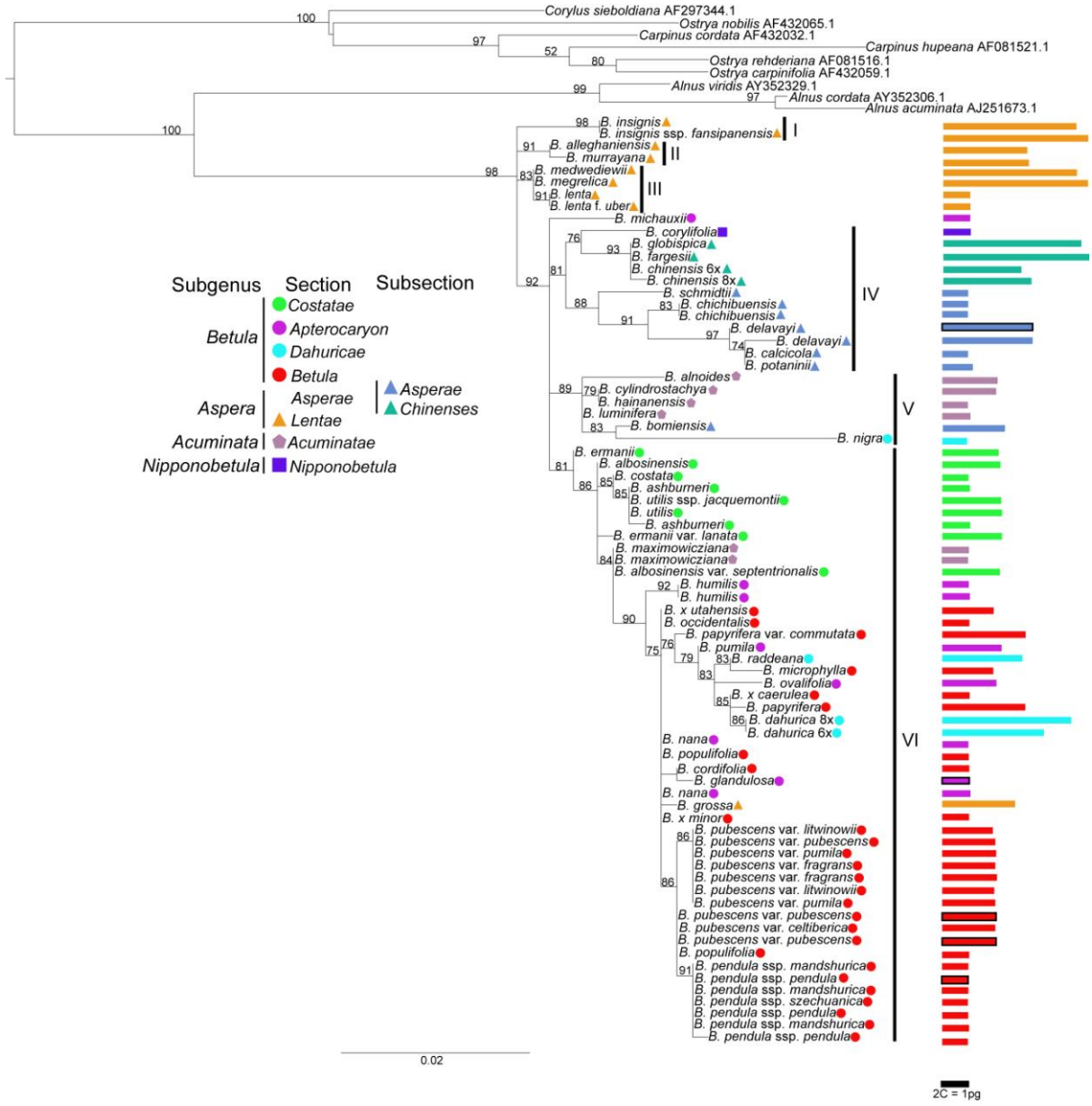
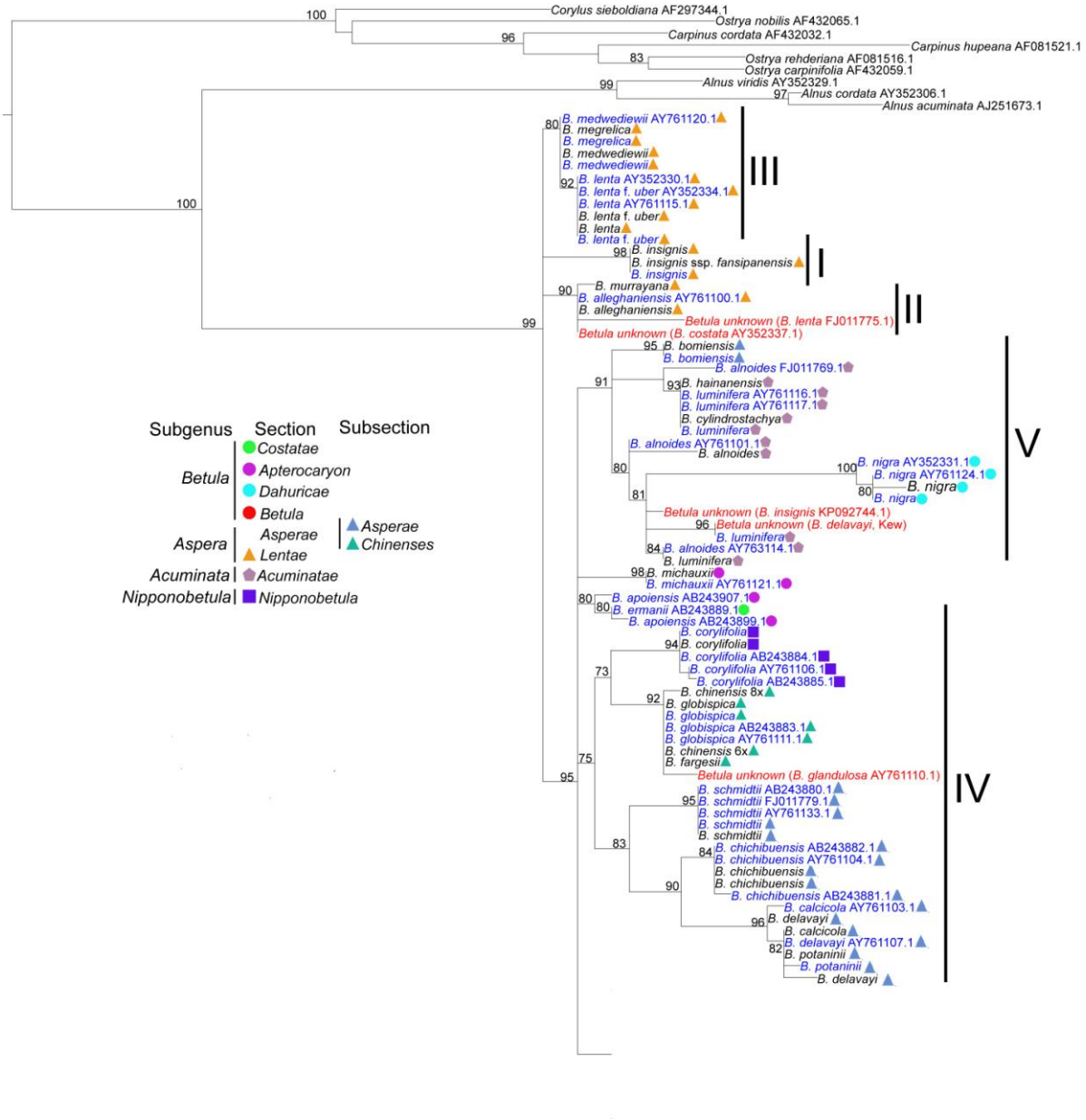


Figure 4.1 Maximum Likelihood analysis of verified *Betula* L. species using ITS sequences. Species were classified according to Ashburner and McAllister (2013). Values are bootstrap percentages which are above 50%. The bars on the right side indicate the genome size of this species with colors corresponding to the taxonomy. Bars with black outlines indicate a tentative genome size of the individual.

The aligned ITS data matrix for all accessions contains 233 ITS sequences and 622 characters of which 188 characters are variable and 132 informative. The phylogeny of all samples reveals a similar overall topology as that of the phylogeny based only on verified samples. However, 24 (16%) of the 148 unverified samples have unexpected phylogenetic positions. Of these 24, half were downloaded from GenBank and half were sequenced from samples collected from botanic gardens. Putative *B. lenta* (Genbank accession FJ011775.1) and *B. costata* (Genbank accession AY352337.1) appear within Clade II, whereas verified accession for these species are in Clade III

and Clade VI, respectively. One putative accession of *B. glandulosa* (Genbank accession AY761110.1) appeared within Clade IV, a clade of species mainly of subsection *Chinenses* whereas another three unverified *B. glandulosa* accessions (Genbank accession FJ011774.1, RBG Kew DNA bank ID: 19950 and Helsinki Botanic Garden accession 1986-0630) are placed in Clade VI. One accession of *B. insignis* (Genbank accession KP092744) and *B. delavayi* (RBG Kew accession 1993-3034) are unexpectedly placed within Clade V whereas the verified samples for these species are in Clade I and Clade IV, respectively. An accession of putative *B. dahurica* (Genbank accession FJ011773) and one of putative *B. skvortsovii* are clustered with *B. utilis* in Clade VI and one accession of *B. chinensis* (Genbank accession AY761105.1) is clustered with seven accessions of *B. dahurica* in Clade VI (Fig. 4.2). All the remaining 12 unverified accessions found unexpectedly in Clade VI cluster with *B. pubescens*, *B. pendula* and their subspecies/varieties (Fig. 4.2).


Corylus sieboldiana AF297344.1
Ostrya nobilis AF432065.1
Carpinus cordata AF432032.1
Carpinus hupeana AF081521.1
Ostrya rehderiana AF081516.1
Ostrya carpinifolia AF432059.1
Alnus viridis AY352329.1
Alnus cordata AY352306.1
Alnus acuminata AJ251673.1
B. medwediewii AY761120.1
B. megrelica
B. megrelica
B. medwediewii
B. medwediewii
B. lenta AY352330.1
B. lenta f. uber AY352334.1
B. lenta AY761115.1
B. lenta f. uber
B. lenta
B. lenta f. uber
B. insignis
B. insignis ssp. fansipanensis
B. insignis
B. murrayana
B. alleghaniensis AY761100.1
B. alleghaniensis
Betula unknown (B. lenta FJ011775.1)
Betula unknown (B. costata AY352337.1)
B. bomiensis
B. bomiensis
B. alnoides FJ011769.1
B. hainanensis
B. luminifera AY761116.1
B. luminifera AY761117.1
B. cylindrostachya
B. luminifera
B. alnoides AY761101.1
B. alnoides
B. nigra AY352331.1
B. nigra AY761124.1
B. nigra
B. nigra
Betula unknown (B. insignis KP092744.1)
Betula unknown (B. delavayi, Kew)
B. luminifera
B. alnoides AY763114.1
B. luminifera
B. michauxii
B. michauxii AY761121.1
B. apoiensis AB243907.1
B. ermanii AB243889.1
B. apoiensis AB243899.1
B. corylifolia
B. corylifolia
B. corylifolia AB243884.1
B. corylifolia AY761106.1
B. corylifolia AB243885.1
B. chinensis 8x
B. globispica
B. globispica
B. globispica AB243883.1
B. globispica AY761111.1
B. chinensis 6x
B. fargesii
Betula unknown (B. glandulosa AY761110.1)
B. schmidtii AB243880.1
B. schmidtii FJ011779.1
B. schmidtii AY761133.1
B. schmidtii
B. schmidtii
B. chichibuensis AB243882.1
B. chichibuensis AY761104.1
B. chichibuensis
B. chichibuensis
B. chichibuensis AB243881.1
B. calcicola AY761103.1
B. delavayi
B. calcicola
B. delavayi AY761107.1
B. potaninii
B. potaninii
B. delavayi
I
II
III
IV
V
Subgenus
Section
Subsection
Betula
Costatae
Apterocaryon
Dahuricae
Betula
Aspera
Asperae
Lentae
Asperae
Chinenses
Acuminata
Acuminatae
Nipponobetula
Nipponobetula

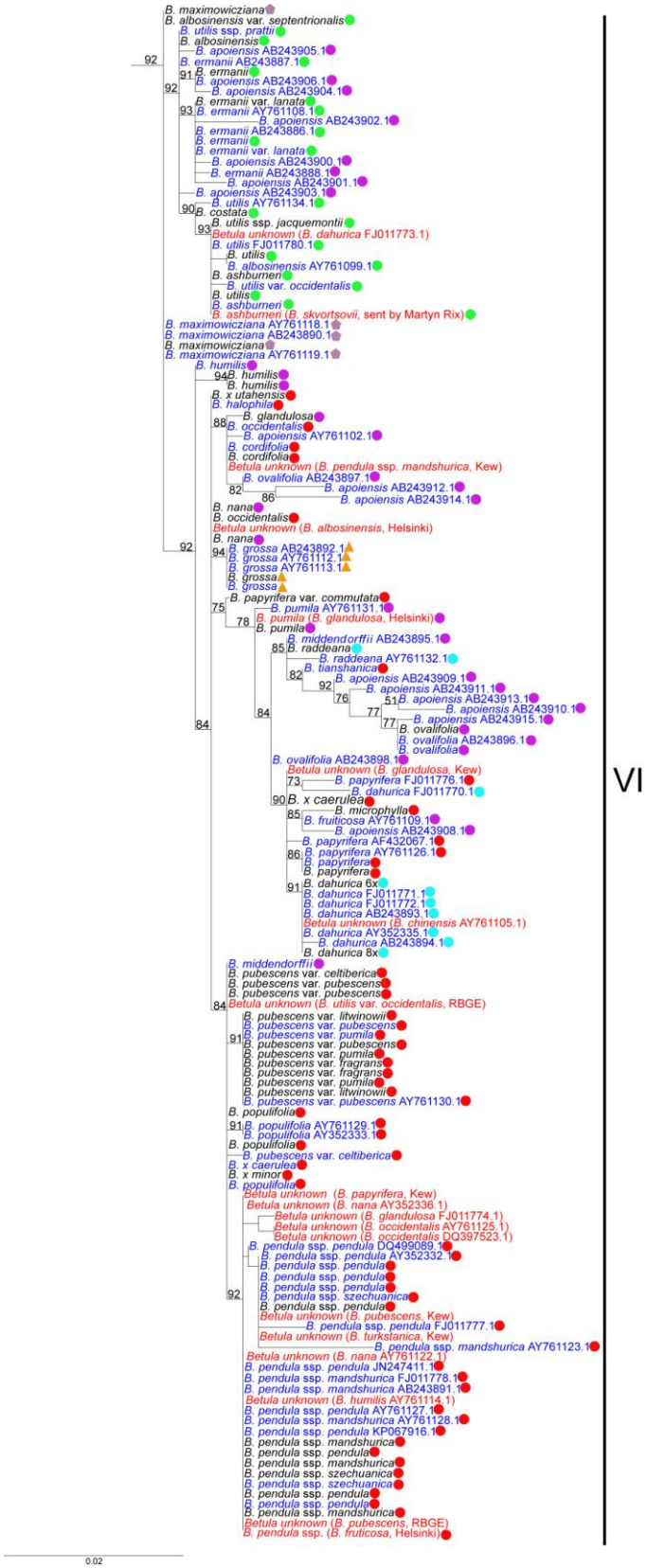


Figure 4.2 Maximum Likelihood analysis of all obtained *Betula* L. samples using ITS sequences. Species were classified according to Ashburner and McAllister (2013). Values are bootstrap percentages which are above 50%. Marked in red, blue and black represent potentially misidentified species, potentially correctly identified species and 'verified' species, respectively. Included in parentheses are the original label of potentially misidentified species and their sources; the suggested correct identification of these species is placed before the parentheses. If unknown, *Betula unknown* was used instead.

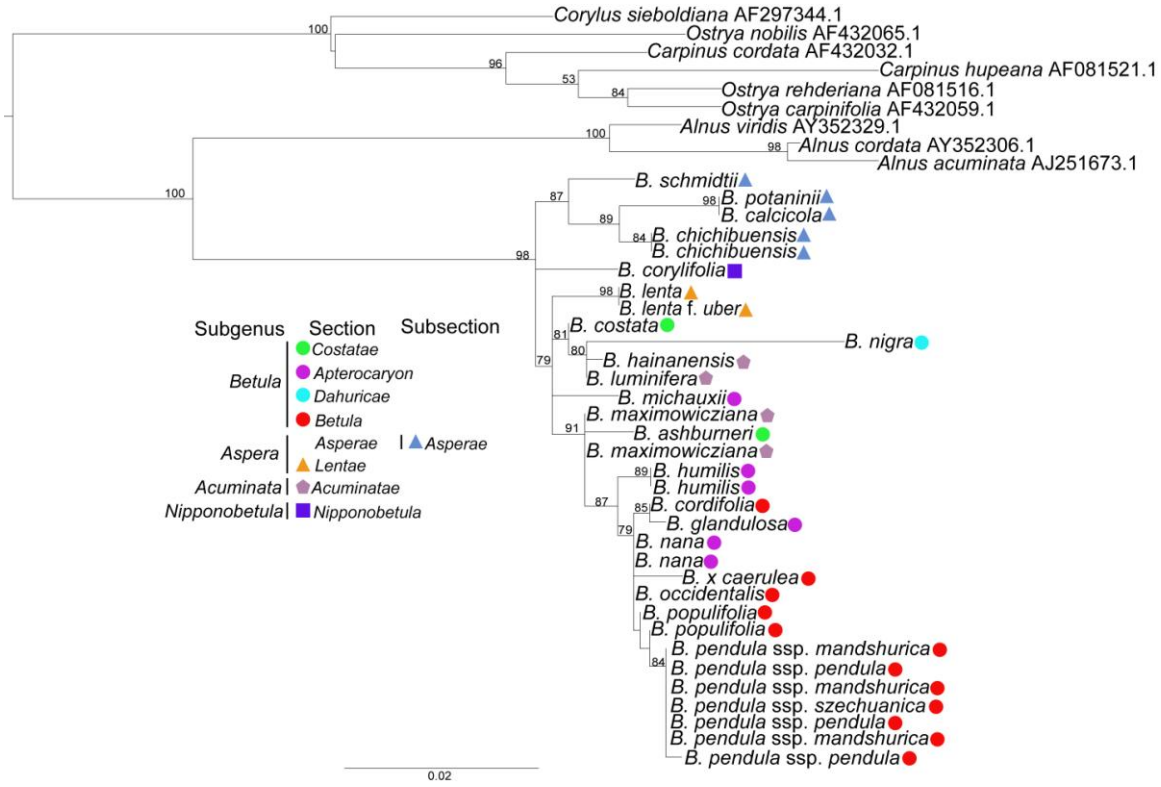


Figure 4.3 Phylogenetic tree from the maximum likelihood analysis of verified *Betula* diploids using ITS. Species were classified according to Ashburner and McAllister (2013). Values above branches are bootstrap percentages of ≥ 50%.

The phylogeny of diploid Betula *accessions*

Betula diploids reveal similar phylogenetic positions as when polyploids were included with a few exceptions: *B. corylifolia* is in a polytomy with subsection *Asperae*; *B. lenta* and *B. lenta* f. *uber* are sister to species of subgenera *Betula* and *Acuminata* whereas *B. costata* clusters with subgenus *Acuminata* (Fig. 4.3).

Genome sizes

We found the 2C genome sizes of *Betula* species to range from 0.88 pg in *B. nigra* to 5.33 pg in *B. insignis* ssp. *fansipanensis*, thus the 1C-value ranges from 0.44 pg (430 Mbp) to 2.67 pg (2611 Mbp). We found Chinese *B. alnoides* to have a genome size of 1.95 pg indicating it is tetraploid rather than diploid (Fig. 4.1). The fact that *B. alnoides* is tetraploid has been subsequently confirmed by chromosome counting and microsatellite genotyping (pers. comm., Hugh McAllister and Jie Zeng). We found a genome size of 0.91 pg for *B.hainanensis*, indicating for the first time that this recently discovered species is diploid. If all other ploidy levels given in Ashburner and McAllister (2013) are correct, the monoploid genome size of *Betula* (1Cx-value) ranges from 371 Mbp for *B. murrayana* to 616 Mbp for *B. dahurica* (Fig. 4.4). The

monoploid genome size is similar among all diploids except for *B. potaninii*. Variance in monoploid genome size is greater among polyploid accessions. There is a significant difference in the variance of 1Cx-values among the groups of 2x, 4x, 6x and 8x-12x accessions (Fig. 4.5, $P < 0.05$; treated pairwise, all groups are significantly non-homogenous in their variances except 4x and 6x ($P = 0.15$) and 6x and 8x-12x ($P = 0.38$)). The proportion of polyploid species of this genus is ~0.60, if only species, subspecies/varieties and different cytotypes are included and species having synonyms are treated as one.

There is a significant difference in the average ploidy level between species with narrow ranges and species with medium, widespread and very widespread ranges (Fig. 4.6A, $P < 0.05$) but no significant difference in the average ploidy level for species with medium, widespread and very widespread ranges (Fig. 4.6A, $P > 0.05$). Similar results also hold true for 2C-values (Fig. 4.6B).

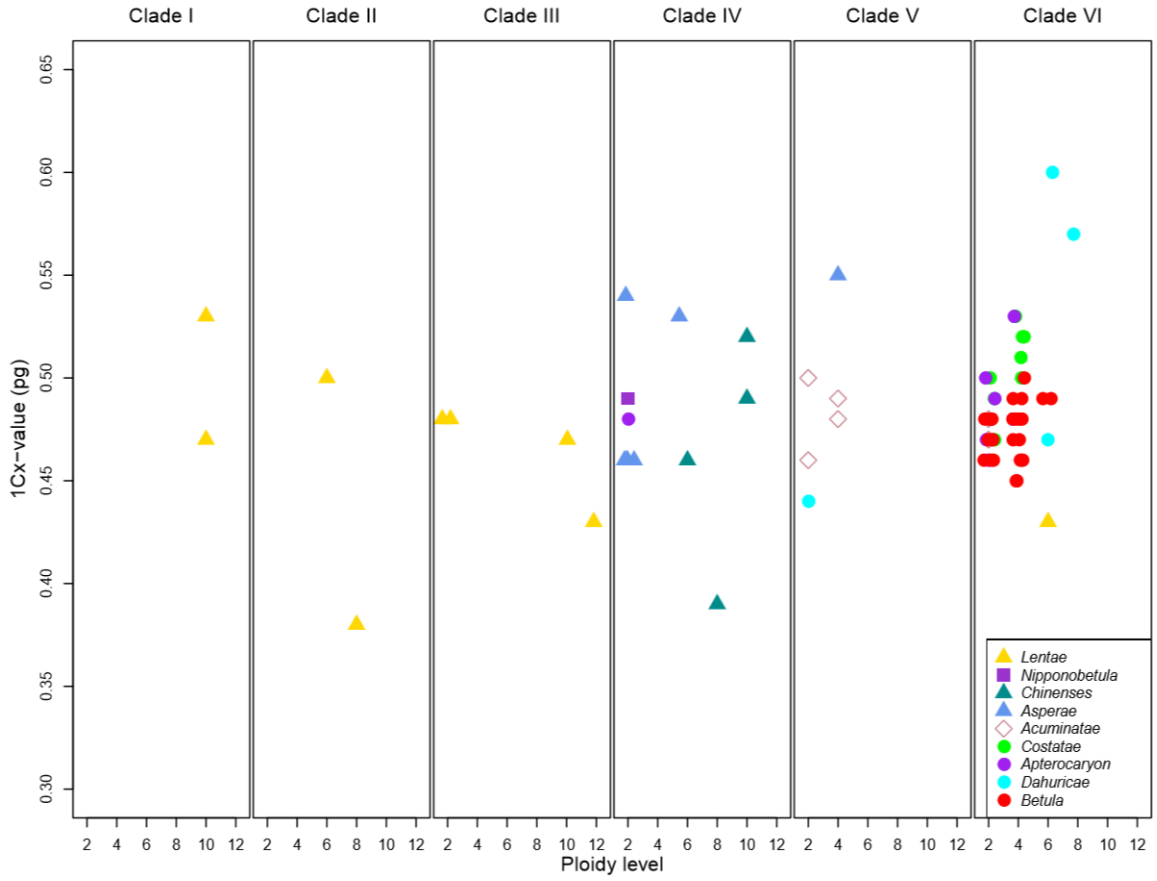


Figure 4.4 The genome size of the basic haplotype (i.e. the 1x value) of *Betula* species and cytotypes measured from verified samples. Ploidal levels were taken from Ashburner and McAllister (2013). Species are depicted in the clades similarly shown in **Figure 4.1**.

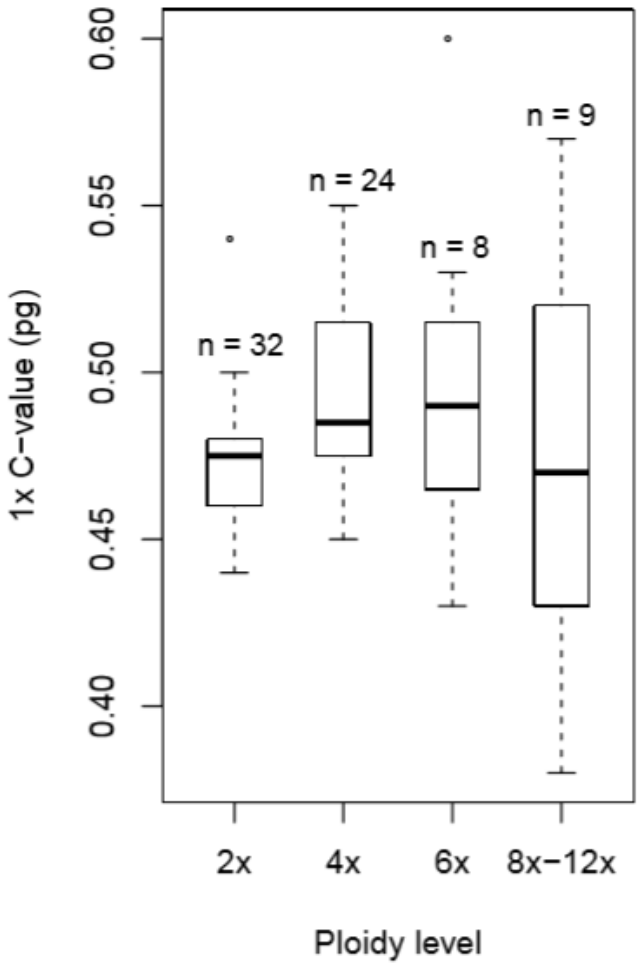


Figure 4.5 The variance of genome size of the basic *Betula* haplotype (1x) of differing ploidy levels: 2x, 4x, 6x and 8x and above. Number of individual in each group is shown above the boxplot.

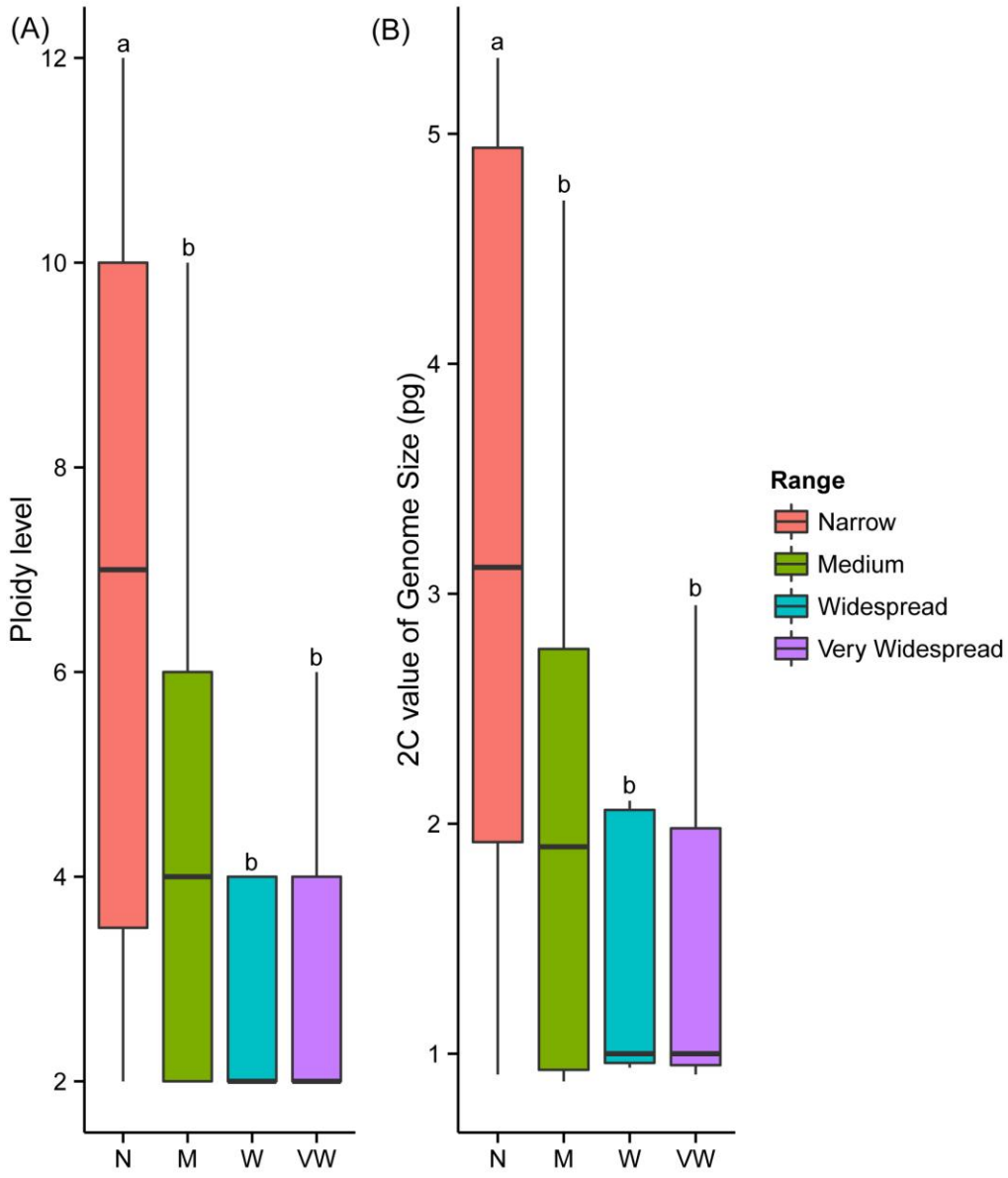


Figure 4.6 The average ploidy level (A) and the mean 2C value of genome size (B) among species of different distribution ranges: narrow, medium, widespread and very widespread, respectively. Letters a and b indicate differences at the significance level at $P \leq 0.05$. There is no significant difference in the average ploidy level or the mean 2C value of genome size if different categories share the same letter.

Discussion

Phylogenetics and taxonomy

Subgenus *Aspera*

Ashburner and McAllister (2013) divided subgenus *Aspera* into two sections: section *Lentae* (from Regel 1865) and section *Asperae*. Our ITS data supports this division, as the majority of species in these two sections fall into distinct ITS clades, athough section *Lentae* is further subdivided into three unresolved clades. The AFLP data of Schenk *et al.* (2008) also agrees with the division of sections *Lentae* and *Asperae*. Ashburner and McAllister (2013) further divided section *Asperae* into subsections *Chinenses* and *Asperae*, which are synonymous with section *Chinenses* and section *Asperae* of Skvortsov (2002), respectively. Our ITS data broadly support this division. Our ITS data do not support Winkler's (1904) combination of sections *Lentae* and *Costatae* of Regel (1865) into subsection *Costatae*. Nor does the data support subgenus *Neurobetula* of De Jong (1993), which consists of species from section *Asperae*, section *Costatae* and section *Dahuricae* of Ashburner and McAllister (2013). In addition, our ITS do not support subgenus *Betulenta* of De Jong (1993) including species such as *B. lenta*, *B. lenta* f. *uber* and *B. globispica* as *B. globispica* is placed in a distinct clade (Fig. 4.1).

The tetraploid species *B. bomiensis*, which Ashburner and McAllister place within section *Asperae* is clustered by ITS into a group of species of subgenus *Acuminata*, but as sister to *B. nigra* which Ashburner and McAllister place in section *Dahuricae.* As Ashburner and McAllister (2013) note, *B. bomiensis* is morphologically similar to *B. potaninii* (section *Asperae*) suggesting that this diploid species may be a parent of *B. bomiensis*. Our genome size data support this hypothesis, in that the monoploid genome size (1Cx) is unusually large for both species (0.54 pg for *B. potaninii* and 0.55 pg for *B. bomiensis*) (Fig. 4.4, Table 4.3). The hypothesis that *B. bomiensis* was formed via hybridisation between *B. potaninii* and a species of subgenus *Acuminata* merits further research with additional genetic loci.

Decaploid species *B. medwediewii* and dodecaploid species *B. megrelica* form a well-supported clade with diploid species *B. lenta* and *B. lenta* f. *uber* (Fig. 4.1 and Fig. 4.2). This suggests that *B. lenta* or its ancestral lineage may have been a parent of these two polyploid species. The morphology of the four species also supports this hypothesis

(Hugh McAllister, unpublished observations). The study of Li *et al.* (2005) found similar result that *B. lenta* and *B. lenta* f. *uber* formed a clade with *B. medwediewii* despite *B. megrelica* was not included. It has previously been suggested (Barnes & Dancik, 1985) that the octoploid species *B. murrayana* is a recent allopolyploid derivative from *B.* x *purpusii*, an inter-subgenus hybrid between *B. alleghaniensis* (8x) and *B. pumila* (4x). We find it to form a clade with *B. alleghaniensis* in the ITS tree supporting this species as one of its parents (Ashburner & McAllister, 2013).

Interestingly, *B. delavayi*, a hexaploid species, clustered with the diploid species *B. calcicola* and *B. potaninii*, indicating that one of these species or their common ancestor could be a parental species of *B. delavayi*. Interestingly, both *B. potaninii* (1Cx = 0.54 pg) and *B. delavayi* (1Cx = 0.53 pg) have an unusually large monoploid genome size, which could be evidence favouring *B. potaninii* as its parental species rather than *B. calcicola* (1Cx = 0.46 pg). Further research is needed to confirm whether other species may also be potential progenitors of *B. delavayi*.

Ashburner and McAllister (2013) place the hexaploid species *B. grossa* in section *Lentae* due to clear morphological similarities, but is not clustered with species of that section by ITS (Fig. 4.1). This is consistent with AFLP data of Schenk *et al.* (2008) and the ITS sequences of Nagamitsu *et al.* (2006). In our case, both *B. grossa* accessions are from different botanic gardens but each shows the same result (Fig. 4.2), making misidentification less likely. The unexpected placement of *B. grossa* into a clade of species of subgenus *Betula* may indicate that one of the progenitors of this polyploid belongs to subgenus *Betula*. It is perhaps an allopolyploid formed from hybridisation with a species of section *Lentae* which it has morphological similarity to, causing McAllister and Ashburner to place it in that section. The ITS sequences from *B. grossa* may be homogenized from one parent (Nagamitsu *et al.*, 2006). This hypothesis for the parentage of *B. grossa* deserves further investigation with a larger number of genetic loci.

Subgenus *Nipponobetula*

Subgenus *Nipponobetula*, which comprises the single species *B. corylifolia*, forms a moderately supported clade (IV) with species of subgenus *Aspera* in this study. This does not support the placement of *B. corylifolia* in section *Costatae* as in Regel (1865), or subsection *Costatae* as in Winkler (1904), or subgenus *Betulenta* as in De Jong (1993). The placement of *B. corylifolia* with subgenus *Aspera* was also indicated in two previous phylogenetic studies (Li *et al.*, 2005; Nagamitsu *et al.*, 2006). However,

we note that *B. corylifolia* is found in an ITS clade within *Aspera* that is composed of the polyploid species *B. chinensis, B. fargesii* and *B. globispica*, and this clade of four species is sister to a clade containing the diploid *Aspera* species, of subsection *Asperae*. We cannot therefore exclude the possibility that *B. corylifolia* is a parental species of allopolyploids *B. chinensis, B. fargesii* and *B. globispica*, through hybridisation with a species from section *Asperae*, and may appear nested in the subgenus *Aspera* as a result. Indeed, in phylogenetic analyses that include only diploid species, *B. corylifolia* is not nested within subgenus *Aspera*, but is in a polytomy with that clade.

Subgenus *Acuminata*

The subgenus *Acuminata* does not form a distinct clade in our ITS phylogenies. Four of its species appear in a clade with *B. nigra*, an outlier from subgenus *Betula* and *B. bomiensis*, an outlier from subgenus *Aspera*. Of these four species, *B. alnoides* and *B. cylindrostachya* are tetraploid and *B. hainanensis* and *B. luminifera* are diploid species, suggesting that one or both of the two diploids or their common ancestor could be parental species of the tetraploids. A fifth species of *Acuminata, B. maximowicziana*, appears in the subgenus *Betula.* A close relationship of *B. maximowicziana* to species of section *Costate* (subgenus *Betula*) is also supported by AFLP markers (Schenk *et al.*, 2008) (though other species of subgenus *Acuminata* were not included in the AFLP study of Schenk *et al.* (2008)). In contrast, the low-copy nuclear gene *NIA* supports the grouping of *B. maximowicziana* with *B. alnoides*, another species of subgenus *Acuminata* (Li *et al.*, 2007), making the phylogenetic position of this species questionable. Two lines of evidence in addition to our ITS results may suggest that *B. maximowicziana* is closely related to species of subgenus *Betula*. First, a crossing experiment apparently showed that fertile hybrids can form between *B. maximowicziana* and *B. pendula* ssp. *mandshurica* (Johnsson, 1945), indicating that no post-zygotic barriers exist; however, this result has not been convincingly reproduced and we thus cannot exclude the possibility that pollen contamination could have occurred. Second, the autumn fruiting and much thicker male catkins of *B. maximowicziana* are distinct from other species of subgenus *Acuminata* (Ashburner & McAllister, 2013). Although the overall appearance and detailed characteristics of *B. maximowicziana* suggest a close relationship with other species of subgenus *Acuminata*, it does stand apart from them in several features, suggesting ancient genetic contribution from another evolutionary line within the genus. If the subgenus *Acuminata* is not monophyletic, the racemose pistillate inflorescence which

characterises it is possibly due to convergent evolution.

Subgenus *Betula*

The majority of the species of the subgenus *Betula* form a single clade, but the four sections of this subgenus have complex relationships in the ITS tree. Section *Costatae* shows a close relationship to section *Betula* and section *Apterocaryon* species are intermixed with section *Betula* (Fig. 4.1 and Fig. 4.2). Species of section *Betula* may have diverged from a lineage of section *Costatae* recently as the reproductive barrier between the two sections is incomplete: hybrids have been created and reported to be fertile, such as *B. pubescens* x *B. ermanii*, *B. pubescens* x *B. albosinensis* and *B. pendula* x *B. ermanii* (Johnsson, 1945). The status of section *Apterocaryon*, containing *B. michauxii* and *B. apoiensis*, *B. nana*, *B. ovalifolia*, *B. fruticosa*, *B. pumila*, *B. humilis* and *B. glandulosa*, defined by dwarf character, is not supported by the ITS tree, which indicates that the dwarf birches are heterogeneous (Fig. 4.1 and Fig. 4.2). This study, together with several other studies (Li *et al.*, 2005; Li *et al.*, 2007; Schenk *et al.*, 2008) suggests that dwarfism is a convergent trait, perhaps due to adaption to cold temperature as evidenced by the existence of bud scales (De Jong, 1993). *Betula nana* shows a closer-relationship with *B. pubescens/B. pendula* than *B. humilis* (Fig. 4.1). Similar result has been indicated by *ADH* (Järvinen *et al.*, 2004) and *NIA* (Li *et al.*, 2007). In addition, the more similar flavonoid profiles of the buds of *B. nana* and *B. pubescens* compared with those between *B. nana* and *B. humilis* (Wollenweber, 1975) suggest a closer-relationship of the former pair than the latter. Surprisingly, *B. michauxii*, a species morphologically almost identical to *B. nana*, is not placed within subgenus *Betula* (Fig. 4.1), which is consistent with the *NIA* phylogeny (Li *et al.*, 2007). Further research is needed to decipher the phylogenetic position of *B. michauxii*.

The taxonomy of the widespread species *B. pendula* and its tetraploid relative *B. pubescens* have been particularly controversial in the past, with several subspecies or varieties of both being described and sometimes classified as independent species. Our analysis (Fig. 4.1 and Fig. 4.2) supports the taxonomic treatment of these two species suggested by Ashburner and McAllister (2013), where taxa within the two species are not given species status. *Betula pubescens* is a tetraploid species; its close relationship with *B. pendula* indicates the possible involvement of *B. pendula* in its formation, as has previously been suggested (Howland *et al.*, 1995). The morphological diversity found within these species is likely due to their wide distribution ranges with morphological variation shaped by overall climatic factors, similar to the variation

found within *B. papyrifera* in N. America (Pyakurel & Wang, 2013). Another factor may be hybridisation and gene flow between *Betula* species in different areas of their distributions.

Within section *Costatae*, *B. costata* forms a well-supported clade with other species of section *Costatae* such as *B. utilis* based on ITS (Fig. 4.1). This supports the inclusion of *B. costata* and *B. utilis* in section *Costatae* (Skvortsov, 2002; Ashburner & McAllister, 2013). Within Clade V, the tetraploid species, *B. alnoides* and *B. cylindrostachya* form an unresolved cluster with the two diploid species, *B. luminifera* and *B. hainanensis*, indicating their common ancestry (Fig. 4.1).

Betula nigra is placed outside the subgenus *Betula* in all of our ITS phylogenies, both with and without unverified samples, and with and without polyploids in the analyses (Fig. 4.1, Fig. 4.2 and Fig. 4.3). In contrast, a phylogenetic study based on *NIA* suggests that it is more closely-related to species of subgenus *Betula* than *B. alnoides* (Li *et al.*, 2007), and morphologically *B. nigra* is most similar to *B. dahurica* (subgenus *Betula*). The phylogenetic position of *B. nigra* needs further research based on multiple loci.

Genome size and ploidy evolution

Different ploidy levels are present in all subgenera and sections of *Betula* except subgenus *Nipponobetula*, indicating several independent occurrences of polyploidy in the evolution of the genus (Järvinen *et al.*, 2004). Only subgenus *Aspera* contains ploidy levels above octoploid (Fig. 4.4, Table 4.2).

The narrow ranges of species of subgenus *Aspera* with high ploidy level (e.g. *B. insignis*, *B. megrelica*, *B. globispica* and *B. fargesii*) may indicate they are of recent origin or have low invasiveness perhaps due to low growth rate, which has been associated with larger genome size (Lavergne *et al.*, 2010; Fridley & Craddock, 2015), or their lack of, or very narrow, seed wings (Ashburner & McAllister, 2013). The narrow distributions of these relatively large genomes may also be influenced by available nutrients, such as nitrogen or phosphorus which may select against plants with large genome sizes (Knight *et al.*, 2005; Leitch & Leitch, 2012; Šmarda *et al.*, 2013), and low temperature, which may influence the rate of cell division (Grime & Mowforth, 1982). On the other hand, these high ploidy level birches occur in areas known to harbor many relictual species, and their small populations may be relicts from larger distributions in the past. In contrast, the most diversified, widespread and 'successful' species are members of subgenus *Betula* with low ploidy levels (such as *B.*

pendula, B. nana and *B. glandulosa*). Hybridisation and adaptive introgression occurr frequently within subgenus *Betula* (Thórsson *et al.*, 2010), which may play an important role in colonisation of new habitats.

Our genome size results agree with published genome sizes for Icelandic birches, *B. nana* and *B. pubescens*, which suggest that no significant genome downsizing has occurred in tetraploid *B. pubescens* (Anamthawat-Jónsson *et al.*, 2010). However, our results for the 2C-value of *B. populifolia* are over twice as large as those measured by Feulgen microdensitometry (Olszewska & Osiecka, 1984). This is unlikely to be simply due to the difference in methodology, as flow cytometry and Feulgen microdensitometry were shown to give congruent measurements for Icelandic birches (Anamthawat-Jónsson *et al.*, 2010). Specimen misidentification is also unlikely to be the cause of the differences, as all of the *Betula* species that we measured have a 2C-value of more than twice the measure of the 2C-value of *B. populifolia* measured by (Olszewska & Osiecka, 1984); perhaps chemical interference (Greilhuber, 2008) is the explanation for their unusual result. We also found the previously reported 2C-value of *B. nigra* at 2.90 pg (Bai *et al.*, 2012) to be large compared to the 2C-value of 0.88 pg for *B. nigra* here, and the specimen measured by Bai *et al.* (2012) has now been identified as *B. alleghaniensis* through checking the voucher specimen (DOB0420) (pers. comm. from Prof. Waller), which is congruent with the 2C-value of 2.97 pg of *B. alleghaniensis* found here (Table 4.2).

We found the monoploid genome size (1Cx-value) for most species of *Betula* to be between 0.42 pg and 0.57 pg. Four outlier species, two with lower 1Cx-values and two with higher 1Cx-values, all have higher ploidy levels: octoploid *B. murrayana* (1Cx = 0.38 pg), octoploid *B. chinensis* (1Cx = 0.39 pg), hexaploid *B. dahurica* (1Cx = 0.60 pg) and octoploid *B. dahurica* (1Cx = 0.57 pg). The chromosome counts of these accessions need to be double-checked, but assuming they are correct, we found a general pattern that the variance of 1Cx genome sizes is greater in the species of *Betula* with higher ploidy levels than it is in the diploid species. This suggests that upsizing or downsizing of the sizes of the genomes is occurring in the polyploid birches, perhaps through loss of genome fragments (Buggs *et al.*, 2009; Buggs *et al.*, 2012a), or proliferation of transposable elements (Bennetzen *et al.*, 2005).

Biogeography

The phylogeography of several species of *Betula* has been extensively studied. In general, widespread species, such as *B. pubescens*/*B. pendula* (Maliouchenko *et al.*,

2007; Thórsson *et al.*, 2010) in Europe and *B. papyrifera*/*B. alleghaniensis* in Northern America (Thomson *et al.*, 2015) show little population subdivision even at large scale, perhaps due to rapid population growth and high levels of gene flow, due to dispersal of pollen and seeds over long distances. In contrast, species likely to have lower dispersal ability such as *B. nana* (Thórsson *et al.*, 2010; Wang *et al.*, 2014a), *B. humilis* (Jadwiszczak *et al.*, 2012) and *B. maximowicziana* (Tsuda & Ide, 2005; Tsuda *et al.*, 2015), reveal a more sub-divided genetic population structure. In addition, geographic barriers in the past and present may play an important role in causing genetic discontinuity (Eidesen *et al.*, 2013).

To our knowledge, biogeographical disjunctions among *Betula* species have only been mentioned in Li *et al.* (2005), based on a smaller sample size. Species of Clade III have disjunct distributions (Ashburner & McAllister 2013), with *B. medwediewii* and *B. megrelica* in Georgia and Turkey, and *B. lenta* in North America. We speculate that their common ancestor may have been continuously distributed over the northern hemisphere. Subsequent climate change may have eliminated it in intervening regions causing geographical disjunctions. In addition, this genus contains three groups with disjunct distributions between North-east Asia and South-west Asia: a common disjunction in groups of related species (Ran *et al.*, 2006). Within subsection *Asperae*, *B. schmidtii* and *B. chichibuensis* occur in NE Asia whereas *B. calcicola*, *B. potaninii* and *B. delavayi* occur only in South-west China. In the clade comprising subsection *Chinenses*, *B. globispica* occurs in North-east Asia, whereas *B. fargesii* occurs in South-west and central China. In the clade comprising *B. costata*, *B. utilis* and *B. ashburneri* (section *Costatae*), the first species occurrs in North-east Asia whereas the latter two are in South-west and central China.

Unexpected phylogenetic positions of unverified accessions

Unexpected phylogenetic signals for a subset of taxa in our phylogeny of all samples led us to re-appraise their identification. The *B. fruticosa* and *B. nana* subsp. *exilis* (synonym *B. glandulosa*) samples from Helsinki Botanic Garden were determined to be a subspecies of *B. pendula* and *B. pumila* respectively based on ITS and morphology (examined by HM). The putative *B. skvortsovii* sample was determined to be *B. ashburneri* based on ITS, morphology (examined by HM) and genome size of 1.00 pg (2C-value). The nesting of two accessions of *B. glandulosa* into a clade including *B. pumila*, whereas the verified *B. glandulosa* was placed into a distinct clade, was probably caused by the misidentification of *B. pumila* as *B. glandulosa* due

to their morphological similarity (Fig. 4.2). Similarly, *B. pendula* is sometimes misidentified as *B. pubescens* and vice versa as there is a continuum of leaf variations between the two (Wang *et al.*, 2014b).

In addition, of the 12 sequences downloaded from GenBank, we think that at least five were possibly misidentified: *B. costata* (AY352337.1), *B. insignis* (KP092744.1), *B. glandulosa* (AY761110.1), *B. dahurica* (FI011773) and *B. chinensis* (AY761105.1). The fact that *B. dahurica* (FI011773) was collected from the Himalaya region is a strong signal of its misidentification because *B. dahurica* is distributed in NE Asia. This species is more likely to be *B. utilis* as *B. utilis* is common in the Himalaya region, and this fits with the ITS data. There are 12 accessions clustered with a clade of *B. pubescens*/*B. pendula*, showing unexpected phylogenetic signals (Fig. 4.2). Besides the one labelled as *B. fruticosa* that is a clear misidentification, the remaining unexpected placements may be caused by hybridisation or gene flow between *B. pubescens*/*B. pendula*, as many species (such as *B. nana*, *B. glandulosa*, *B. humilis*, *B. occidentalis*, *B. turkstanica* and *B. papyrifera*) can hybridise naturally or in cultivation with *B. pubescens*/*B. pendula* (Barnes *et al.*, 1974; Sulkinoja, 1990; Truong *et al.*, 2007; Jadwiszczak *et al.*, 2012; Ashburner & McAllister, 2013) .

Concluding remarks

Phylogenentic analyses of genus *Betula* based on ITS sequences broadly agrees with Ashburner and McAllister's (2013) taxonomical treatment of this genus. This study gives us some new information about the possible origins of some polyploids in the genus, such as *B. alnoides*, *B. chinensis*, *B. delavayi*, *B. medwediewii* and *B. megrelica* but the origins of *B. bomiensis* and *B. grossa* remain ambiguous. The phylogenetic positions of *B. michauxii*, *B. maximowicziana* and *B. nigra* remain questionable. The phylogenetic relationships within genus *Betula* needs to be further addressed using multiple loci and next-generation sequencing methods such as restriction site associated DNA markers, which have been successfully applied to *Betula* species in a pilot study (Wang *et al.*, 2013).

Chapter 5 RAD markers and phylogenomics of *Betula* diploid species

Nian Wang, Jasmin Zohren, Hugh A. McAllister, Richard J. A. Buggs*

Information:

This chapter is formatted to be part of a publication, for which I am going to be the lead author. Jasmin Zohren helped to write an R script to change loci names. All authors will contribute to commenting and proofreading the manuscript.

Summary

Discordance among molecular phylogenetic trees has been frequently observed and attributed to various factors, such as deep coalescence, hybridisation and introgression or horizontal gene transfer. New approaches have been used to construct phylogenetic trees that seek to take into account the history of many separate gene trees and the overall history of species divergence. *Betula* provides an excellent model to study species relationships. In this study, we developed restriction site associated DNA (RAD) markers for most *Betula* species and used a concatenation method and methods based on multiple gene tree summary statistics (STAR and MP-EST) to infer relationships among diploid species of the genus. 587 loci with a minimal length of 500bp were used in the analysis. In addition, we used binary data indicating the presence and absence of RAD tags to infer the species relationships. The results show that *B. corylifolia* and *B. michauxii* should be incorporated into subgenus *Aspera* whereas *B. maximowicziana* should be placed within subgenus *Betula*. *Betula nigra* can be regarded as representing a new subgenus (subgenus *Dahurica*). Based on our phylogenomic approach, we propose a new classification of *Betula* of four subgenera and seven sections (subgenera *Acuminata* [section *Acuminatae*], *Aspera* [sections *Asperae* and *Lentae*], *Dahurica* [section *Dahuricae*], *Betula* [sections *Betula*, *Costatae* and *Maximowiczianae*]). The study shows that RADSeq is suitable for phylogenomic analysis especially when a reference genome is available (albeit fragmented). This dataset will also allow future investigation of the parental origins of polyploid species, which is not included in this study due to time constraints imposed by the length of my scholarship.

Introduction

It is recognised that phylogenetic gene trees are not the same as species trees (Pamilo & Nei, 1988; Nichols, 2001) due to various factors, such as incomplete lineage sorting, hybridisation and introgression of genes, or horizontal gene transfer (Degnan & Rosenberg, 2009). Phylogenomic analysis provides an excellent approach to investigate and compare different phylogenetic gene trees (Maddison, 1997), to detect genes which reflect speciation events, and to seek to build species trees based on a genome-wide dataset. Distantly-related species may appear as sisters in some gene trees due to convergent evolution. Investigating such genes based on phylogenomic approaches provides a window on the past evolution of species' genomes. Two classes of methods are often used to infer the species tree from multi-locus data: concatenation methods and coalescence-based methods. Phylogenetic analysis based on a concatenated supermatrix, assuming all loci have the same evolutionary history, can be problematic (Degnan & Rosenberg, 2006; Kubatko & Degnan, 2007). In recent years coalescence-based methods have gained popularity in inferring species trees, taking into account the variation of individual gene trees (Maddison & Knowles, 2006; Liu, 2008; Liu *et al.*, 2008; Kubatko *et al.*, 2009).

The rapid advance of next-generation sequencing (NGS) technologies has allowed the generation of millions of reads at a much lower cost and at a high speed compared with traditional Sanger sequencing. NGS has been increasingly applied to non-model species, which have complex genomes and have usually undergone several rounds of duplication events (Jiao *et al.*, 2011). Restriction site associated DNA sequencing (RADSeq) uses NGS methods to sequence the flanking regions of restriction enzyme cutting sites. It has been successfully used in various fields, such as population genomics (Hohenlohe *et al.*, 2010), marker development (Barchi *et al.*, 2011; Etter *et al.*, 2011), phylogenetic reconstruction (Rubin *et al.*, 2012; Cruaud *et al.*, 2014) and phylogeography (Emerson *et al.*, 2010). Most phylogenetic studies based on RADSeq use short reads and a supermatrix approach (Hipp *et al.*, 2014; Pante *et al.*, 2015). In the present study, we develop >500bp RAD tags for nearly all described *Betula* species for phylogenomic analysis, following RADSeq analysis of *B. nana* and *B. pubescens* in a pilot study (Wang *et al.*, 2013).

Betula provides an excellent model to study its species relationships and to infer the parental origins of polyploid species. It is a genus of Betulaceae and is sister to *Alnus* (Chen *et al.*, 1999). It includes ~60 species with ranges across the northern hemisphere and is of high ecological and economic importance (Ashburner & McAllister, 2013). Some species are widely used in horticulture and forestry whereas other species are important trees for timber production. Despite the fact that some species are widespread and invasive, some have restricted distributions and hence have been listed as endangered in the IUCN Red List (Shaw *et al.*, 2014). The genome size (2C-value) of *Betula* species ranges from 0.88 pg to 5.33 pg (Wang *et al.*, unpublished data). Polyploidy is common within *Betula* with ploidy level ranging from diploid to dodecaploid and with the highest chromosome number 2n = 168 (Ashburner & McAllister, 2013). Some species contain more than one cytotype, such as *B. chinensis* (6x and 8x) and *B. dahurica* (6x and 8x) (Ashburner & McAllister, 2013), which could result from interspecific hybridisation.

Ashburner and McAllister (2013) classified *Betula* into four subgenera and eight sections: subgenera *Acuminata* (section *Acuminatae*), *Aspera* (sections *Asperae* and *Lentae*), *Betula* (sections *Apterocaryon, Betula*, *Costatae* and *Dahuricae*) and *Nipponobetula* (section *Nipponobetula*). This classification is similar to that of Skvortsov (2002) but placed *Acuminata* as a subgenus rather than a section of subgenus *Betula*. So far, this is the most comprehensive monograph of *Betula* and provides detailed information on its cultivation, biogeography and identification. In addition, it reports the chromosome numbers of nearly all described species and describes the distribution and the morphology of each species (Ashburner & McAllister, 2013).

In the past decade, molecular phylogenies of *Betula* species using nuclear markers *ITS*, *NIA*, *ADH*, chloroplast gene *matK* and AFLP have been conducted to evaluate classifications proposed by Regel (1865), Winkler (1904), De Jong (1993), Skvortsov (2002) and Ashbuner and McAllister (2013) (Järvinen *et al.*, 2004; Li *et al.*, 2005; Li *et al.*, 2007; Schenk *et al.*, 2008). Phylogenetic trees based on these molecular markers are partially inconsistent and all contradict to classifications based on morphological characters. A recent study using verified *Betula* species shows that phylogenetic positions of a few species are questionable, such as *B. corylifolia*, *B. maxmowicziana*, *B. michauxii* and *B. nigra*, which merits further research (Chapter 5). The discordance between phylogeny and morphology-based classifications of *Betula* species is usually

attributed to introgressive hybridisation, morphological convergence and the occurrence of allopolyploidy (Järvinen *et al.*, 2004; Nagamitsu *et al.*, 2006). Indeed, hybridisation within *Betula* has been extensively studied based on morphological characters, cytogenetics, genome size analysis, hand-cross pollination and molecular markers (Johnsson, 1945; Anamthawat-Jónsson & Tómasson, 1990; Anamthawat-Jónsson & Thórsson, 2003; Anamthawat-Jónsson *et al.*, 2010; Wang *et al.*, 2014a). Extensive gene flow has blurred species boundaries and caused taxonomic confusion (Wang *et al.*, 2014b). Also, hybridisation possibly has resulted in ploidy level variation within species. Polyploidy is common within *Betula*, with nearly 60% of its species are polyploids (see Chapter 4). *Betula pendula* has been hypothesised to be involved in the formation of tetraploid *B. pubescens* and hexaploid *B. papyrifera* (Howland *et al.*, 1995; Järvinen *et al.*, 2004). However, the hypothesis remains untested. Hence, RAD markers developed here provide a basis for future research on the parental origins of polyploid species.

The specific aims in the present study are to develop >500bp RAD markers for *Betula* species and to infer phylogenetic relationships of *Betula* diploid species using RAD markers. The recently sequenced whole genome of *B. nana* provides a good reference for mapping.

Materials and Methods

Sample collection

Samples were obtained from the Stonelane Gardens (SL hereafter), the Ness Gardens (N hereafter), the Royal Botanic Gardens Edinburgh (RBGE) or collected by the research group (Table 5.1). The genome size of most of these taxa has been obtained (see Chapter 4). Three species which can serve as outgroups were also included for RADSeq and these are: *Corylus avellana*, *Alnus inokumae* and *A. orientalis*.

Genome size analysis

We measured the genome size of *C. avellana*, *A. inokumae* and *A. orientalis*. Cambial tissue was co-chopped with internal standards: *Solanum lycopersicum* L. "Stupiké polní rané" (Doležel *et al.*, 1998) or *Pisum sativum* L. "Minerva Maple" (Bennett & Smith, 1991) in 1ml Extraction Buffer (Cystain PI absolute P, Partec GMBH) and then filtered into a tube containing 2.0 ml Staining Solution (Cystain PI absolute P, Partec GMBH) with 12 µl propidium iodide (PI). Samples were incubated at room temperature in the dark for c. 30 min. Three to five replicates were analyzed per sample; for each replicate, over 5000 nuclei were measured using a Partec CyFlow Space flow cytometer (Partec, GmbH, Germany) fitted with a 100-mW green solid-state laser (Cobolt Samba; Cobolt, Sweden). The resulting histograms were analyzed with the Flow-Max software (v. 2.4, Partec GmbH).

DNA extraction, RAD library preparation and Illumina sequencing

Genomic DNA was isolated from silica-dried cambial tissue or leaves following a modified 2× CTAB (cetyltrimethylammonium bromide) protocol (Wang *et al.*, 2013). The isolated DNA was assessed with 1.0% agarose gels and measured with a Qubit 2.0 Fluorometer (Invitrogen, Life technologies) using Broad-range assay reagents. The DNA was diluted to a final concentration of ~30 ng/µl for subsequent use.

RAD libraries were prepared following the protocol of Etter *et al.* (2011) with slight modifications (Fig. 5.1A). Briefly, 0.5 ~ 1.0 µg of genomic DNA for each sample was softened by heating at 65°C for 2~3 hours prior to digestion with PstI (New England Biolabs, UK). This enzyme has a 6bp recognition site and leaves a 4bp overhang. Digestion was followed by ligation of barcoded P1 adapters. Ligated DNA was sheared using a Bioruptor (KBiosciences, UK) instrument in 1.5 mL tubes (high intensity, duration 30 s followed by a 30 s pause which was repeated eight times).

Sheared fragments evenly distributed between 100bp and 1500bp and the size of ~600bp was selected using Agencourt AMPure XP Beads (NEB) following a protocol of double-size selection. Briefly, use a ratio of bead:DNA solution of 0.55 to remove large fragments followed by a second size selection by mixing the supernatant from the previous one with 5 ul concentrated beads (20 µl beads with15 µl supernatant removing). After end-repair and A-tailing, the size-selected DNA was ligated to P2 adapters (400 nm) and PCR amplified. PCR amplification was carried out in 25 µL reactions consisting of 0.46 vol ddH_2O and template DNA (4-5 ng), 0.5 vol 2 × Phusion Master Mix (New England Biolabs), and 0.04 vol P1 and P2 amplification primers (10 nm stock), using the following cycling parameters: 98°C for 30 s followed by 12 cycles of 98°C for 10 s and 72°C for 60 s. Three or four independent PCR replicates were conducted for each sample to achieve enough amount of the library. The final library was quantified using the Bioanalyzer and Qubit and normalised prior to sequencing. The quantified library was sequenced on a MiSeq machine using MiSeq Reagent Kit v3 (Illumina) at the Genome Centre of Queen Mary University of London.

RAD data trimming and demultiplexing

The raw data were trimmed using Trimmomatic (Bolger *et al.*, 2014) in paired end mode with the following steps. First, we used LEADING and TRAILING steps to remove bases from the ends of a read if the quality is below 20. Then we performed a SLIDINGWINDOW step with a window size of 1 and a required quality of 20. Finally, we used a MINLENGTH step to discard reads shorter than 100bp. FastQC was used to check various parameters of sequence quality in both raw and trimmed datasets (Andrews, 2014). The trimmed data were demultiplexed, using the process_radtags module of Stacks (Catchen *et al.*, 2013).

Mapping to B. platyphylla genome

The whole genome sequence of *B. platyphylla* has been assembled at chromosomal level (unpublished data from Chinese collaborators), which serves as a reference for mapping. A reference helps to separate homologous loci from paralogous loci (Wang *et al.*, 2013), and to anchor reads with adjacent restriction cutting site (Fig. 5.1B). Mapping of trimmed reads for each sample was conducted in the CLC Genomics Workbench v. 8. Default parameters with a similarity value of at least 0.8 and a fraction value of at least 0.5 were applied. Reads with unspecific match were ignored and any regions with coverage below two were removed. The consensus sequence with a minimal length of 500bp was created for each sample.

Sequence alignment and trimming

We used all 27 successfully sequenced diploid *Betula* samples and one outgroup species (*A. inokuame*) for subsequent analysis. *Betula glandulosa* was excluded in this analysis because the number of mapped loci is only a few hundreds. We chose only one outgroup species to increase the number of loci present in most samples. Loci present in *A. inokumae* and at least 24 *Betula* samples were retrieved and aligned using Mafft v.6.903 (Katoh *et al.*, 2005) with default parameters. Aligned sequences were trimmed using trimAl v1.2rev59 (Capella-Gutierrez *et al.*, 2009); missing data present in 15% or above of taxa were removed (-gt 0.85) for alignment with no missing taxa. As for alignments with missing taxa, any column with missing data in 18% or above of taxa were removed (-gt 0.82). Using a custom Perl script, aligned reads in fasta format were converted to phy format prior to phylogenetic analysis.

Phylogenetic analysis of the concatenated loci

587 loci were concatenated, of these, 73 have no missing taxa and 146, 168 and 200 have one, two or three missing *Betula* taxa, respectively. The concatenated matrix was analysed with maximum likelihood (ML) in RAxML 7.2.8 using GTR+ Γ model with the gamma distribution of rates among sites. A rapid bootstrap analysis with 100 replicates combined with 10 searches for the optimal tree was conducted.

Species tree inference

Individual gene trees for each of the 587 loci were estimated using the maximum-likelihood method (ML) in RAxML 7.2.8 (Stamatakis, 2006) and rooted by *A. inokuame*. A rapid bootstrap analysis with 100 bootstraps and 10 searches was performed for each of the 587 loci under a GTR + Γ model. These bootstrapped gene trees were used to infer the species tree based on two coalescent methods: species tree estimation using average ranks of coalescence (STAR) (Liu *et al.*, 2009a) and maximum pseudo-likelihood estimation of species trees (MP-EST) (Liu *et al.*, 2010) given the distinct topologies among genes trees.

The STAR analyses were conducted using R package ‘phybase’ with the neighborjoining algorithm on a matrix of ranks of taxon pairs in the estimated gene trees. A STAR tree was constructed from each multilocus pseudoreplicate, and a majority rule consensus STAR tree was then built from the 100 replicates in Phylip v.3.695 (Felsenstein, 1989). MP-EST analysis for the 587 loci was conducted in STRAW (Shaw *et al.*, 2013).

Phylogenetic analysis based on the presence/absence of RAD tags

Mapped reads with the minimal length of 500bp was used to construct a matrix of binary absence/presence of RAD tags in each taxon; the absence and presence of RAD tags was coded as '0' and '1', respectively. The matrix was analysed in RAxML 7.2.8. (Stamatakis, 2006), using the GTR + Γ model (general time reversible model) with the gamma distribution of rates among sites. A rapid bootstrap analysis with 100 bootstraps and 10 searches was performed to search the optimal tree.

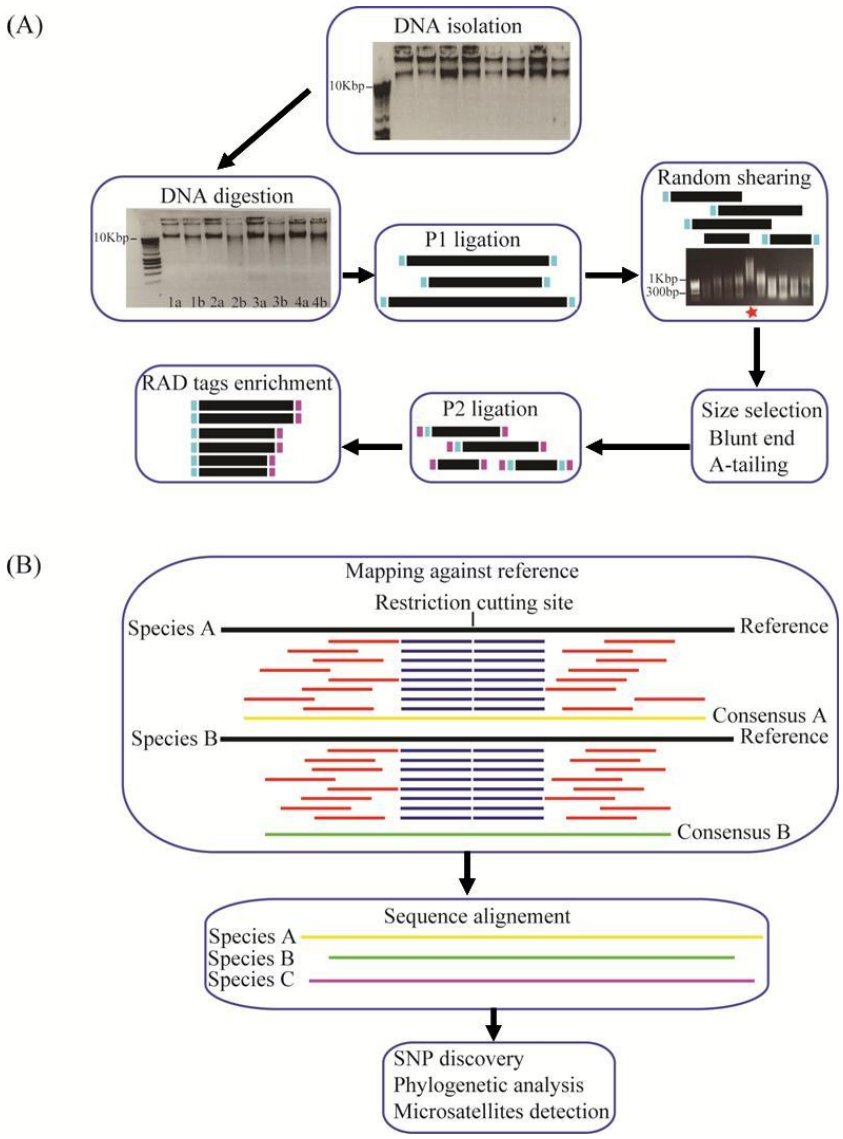


Figure 5.1 Procedures for RAD library preparation (A) and subsequent reads mapping to a reference (B). 1a, 2a, 3a and 4a represent genomic DNA whereas 1b, 2b, 3b and 4b represent the corresponding digested products. A red star indicates sheared fragments with a skewed size distribution.

Table 5.1 Detailed information of the taxa used for restriction site associated DNA sequencing (RADSeq).

Species	Botanic Garden[1]	Section/Subgenus[2]	No. of loci[3]	Total reads[4]	Mapping percentage (%)
B. bomiensis P.C.Li	N	*Asperae/Aspera*	15058	1,554,554	77.29
B. calcicola (W.W.Sm.) P.C.Li	N	*Asperae/Aspera*	19467	1,322,264	74.52
B. chichibuensis Hara	SL	*Asperae/Aspera*	19899	1,669,656	76.57
B. chinensis Maxim. 6x	N	*Asperae/Aspera*	16564	1,263,784	79.05
B. chinensis Maxim. 8x	N	*Asperae/Aspera*	24059	2,229,454	78.63
B. delavayi Franch.	SL	*Asperae/Aspera*	17579	1,482,446	75.42
B. fargesis (Franchet) P. C. Li.	N	*Asperae/Aspera*	24335	2,938,058	69.07
B. globispica Shirai	N	*Asperae/Aspera*	24605	3,270,984	74.58
B. potaninii Batalin	N	*Asperae/Aspera*	18255	1,287,490	74.44
B. schmdittii Regel	N	*Asperae/Aspera*	14906	2,089,830	77.39
B. alleghanensis Britton	N	*Lentae/Aspera*	19856	1,618,440	78.26
B. grossa Siebold & Zucc.	SL	*Lentae/Aspera*	31525	4,948,008	76.06
B. insignis Franch.	SL	*Lentae/Aspera*	25975	2,539,852	80.01
B. lenta L.	SL	*Lentae/Aspera*	19879	1,272,506	75.5
B. lenta f. *uber* (Ashe) Fernald	SL	*Lentae/Aspera*	18303	1,223,142	76.7
B. medwediewii Regel	N	*Lentae/Aspera*	25678	2,363,800	80.24
B. megrelica D. Sosn.	N	*Lentae/Aspera*	49251	7,010,790	80.48
B. murrayana B. V. Barnes & Dancik	N	*Lentae/Aspera*	25534	2,979,142	74.37
B. alnoides Buchanan-Hamilton ex D. Don	&	*Acuminatae/Acuminata*	16840	1,513,890	80.12
B. cylindrostarchy Lindl. ex Wall	SL	*Acuminatae/Acuminata*	14209	1,365,292	78.58
B. hainanensis J. Zeng, B.Q. Ren, J.Y. Zhu & Z.D. Chen	&	*Acuminatae/Acuminata*	16253	1,065,196	79.18
B. hainanensis J. Zeng, B.Q. Ren, J.Y. Zhu & Z.D. Chen	&	*Acuminatae/Acuminata*	15530	1,480,662	72.09
B. luminifera H.Winkl.	RBGE	*Acuminatae/Acuminata*	23027	1,386.43	79.43
B. maximowicziana Regel	N	*Acuminatae/Acuminata*	22402	2,357,348	76.37
B. humilis Schrank	N	*Apterocaryon/Betula*	19389	2,367,672	78.17
B. michauxii Spach	N	*Apterocaryon/Betula*	17745	1,259,206	80.06
B. nana L.	&	*Apterocaryon/Betula*	18960	1,348,154	75.46
B. ovalifolia Ruprecht	SL	*Apterocaryon/Betula*	16567	1,632,526	78.99

B. pumila L.	SL	*Apterocaryon/Betula*	21409	2,529,202	63.94
B. cordifolia Regel	N	*Betula/Betula*	18357	1,137,294	75.03
B. halophila Ching	&	*Betula/Betula*	11537	841,156	79.79
B. microphylla Bunge	N	*Betula/Betula*	18710	2,613,778	79.3
B. occidentalis Hooker	SL	*Betula/Betula*	21011	1,403,626	76.2
B. populifolia Marshall	SL	*Betula/Betula*	21811	2,560,486	78.6
B. pendula Roth ssp. *mandshurica* (Reg.) Nakai	N	*Betula/Betula*	21404	1,324,336	77.13
B. pendula Roth ssp. *mandshurica* (Reg.) Nakai	N	*Betula/Betula*	24468	1,755,416	79.2
B. pendula Roth ssp. *szechuanica* Ashburner & McAll.	SL	*Betula/Betula*	22506	1,543,016	78.91
B. pendula Roth ssp. *pendula* Roth	SL	*Betula/Betula*	23408	1,336,526	78.61
B. pendula Roth ssp. *pendula* Roth	&	*Betula/Betula*	20594	1,626,372	80.56
B. pendula Roth ssp. *pendula* Roth	SL	*Betula/Betula*	24042	1,370,060	76.86
B. pubescens Ehrh. var. *celtiberica* Rivas Mart.	SL	*Betula/Betula*	24049	1,915,178	80.92
B. pubescens Ehrh. var. *fragrans* Ashburner & McAll.	SL	*Betula/Betula*	21723	1,528,340	80.15
B. pubescens Ehrh. var. *litiwinowii* Ashburner & McAll.	SL	*Betula/Betula*	20513	1,317,598	78.63
B. pubescens Ehrh. var. *pubescens*	SL	*Betula/Betula*	14725	1,449,548	79.24
B. pubescens Ehrh. var. *pumila* (L.) Govaerts	SL	*Betula/Betula*	18031	1,476,080	80.51
B. papyrifera Marshall	SL	*Betula/Betula*	22795	1,808,098	78.68
B. papyrifera Marshall var. *commutata* Regel	SL	*Betula/Betula*	24976	1,999,870	80.15
B. tianshanica Rupr.	RBGE	*Betula/Betula*	44555	6,241,386	81.43
B. albosinensis Burkill	SL	*Costatae/Betula*	22281	1,692,098	81.01
B. albosinensis Burkill var. *septentrionalis* C. K. Schneider	SL	*Costatae/Betula*	16568	1,702,006	78.82
B. ashburneri McAllister & Rushforth	SL	*Costatae/Betula*	19704	1,558,150	75.94
B. ashburneri McAllister & Rushforth	&	*Costatae/Betula*	19572	1,168,878	80.71
B. costata Trautv.	N	*Costatae/Betula*	19190	1,210,112	76.8
B. ermanii Cham.	SL	*Costatae/Betula*	21748	1,776,770	78.69
B. ermanii var. *lanata* Regel	SL	*Costatae/Betula*	23586	2,068,018	78.42
B. utilis D.Don var. *occidentalis* Ashburner & A.D.Schill.	RBGE	*Costatae/Betula*	28863	2,368,010	80.7
B. utilis D.Don var. *prattii* Burkill	N	*Costatae/Betula*	22350	1,955,356	80.8
B. utilis D.Don	N	*Costatae/Betula*	27813	2,138,046	79.32

B. dahurica Pall. 6x	SL	*Dahuricae/Betula*	41521	3,292,436	83.17
B. dahurica Pall. 8x	N	*Dahuricae/Betula*	27071	2,698,200	80.61
B. nigra L.	SL	*Dahuricae/Betula*	19279	1,401,592	76.94
B. raddeana Trautv.	SL	*Dahuricae/Betula*	18937	1,330,848	79.09
B. corylifolia Regel & Maxim	SL	*Nipponobetula/Nipponobetula*	19181	1,352,996	73.97
Alnus inokumae S. Murai & Kusaka	SL		9250	1,370,880	50.32
Alnus orientalis Decne.	SL		10034	1,717,010	53.75
Corylus avellana L.	RBGE		7245	1,364,940	51.7

[1]SL: Stone Lane Gardens; N: Ness Gardens; RBGE: Royal Botanic Garden Edinburgh; [2]Species are classified according to Ashbuner and McAllister (2013); [3]The number of loci which has a minimal length of 500bp. Any regions with coverage below two were removed. [4]Reads have been trimmed and only reads with a length above 100bp were counted.

Results

RAD data description

The number of reads per sample after trimming and filtering ranges from 841,156 in *B. halophila* to 7,010,790 in *B. megrelica* with an averge of 1,942,445. A proportion of 50.32% of reads from *A. inokumae* is mapped to *B. platyphylla* genome whereas this figure is 81.43% for *B. tianshanica* (Table 5.1). The number of mapped loci with a minimal length of 400bp ranges from 10,705 (*C. avellana*) to 29,593 (*B. luminifera*) in diploid species and from 19,085 (*B. cylindrostachy*) to 63,297 (*B. megrelica*) in polyploid species excluding *B. halophila* with an unknown ploidy level. This number of loci with a minimal length of 500bp ranges from 7,245 to 24,468 and from 14,209 to 49,251 in the same diploid and polyploid species, respectively. The number of loci with a minimal length of 600bp ranges from 2,771 (*C. avellana*) to 17,108 (*B. luminifera*) and from 5,873 (*B. pubescens*) to 32,704 (*B. megrelica*) in diploid species and polyploid species, respectively (Fig. 5.2). Seventy-three loci with a minimal length of 500bp are present in all 28 successfully RAD sequenced diploid individuals whereas 587 are present allowing at most three missing *Betula* sequences per locus (Fig. 5.3).

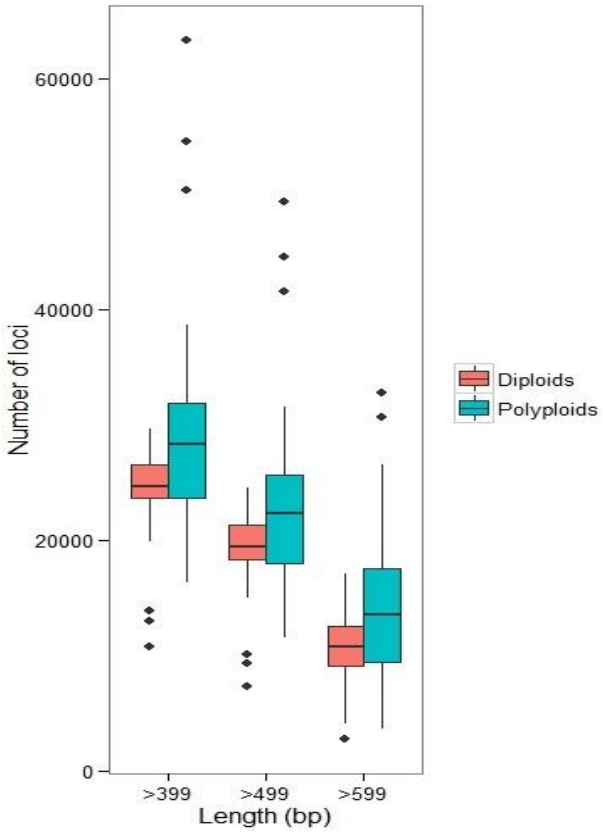


Figure 5.2 Number of loci with a minimal length of 400bp, 500bp and 600bp in 30 diploid taxa and 36 polyploid taxa, respectively. Any regions within loci with coverage below two were removed.

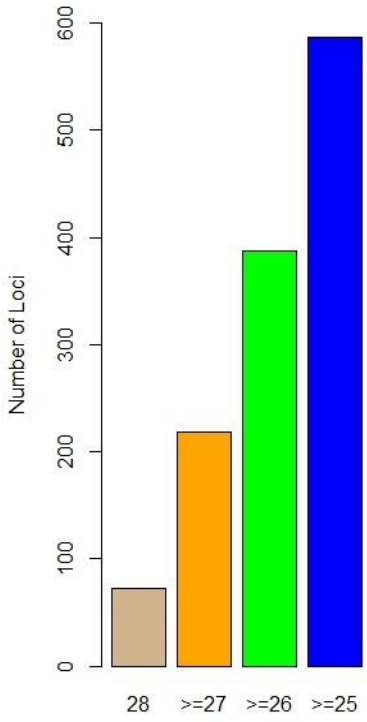


Figure 5.3 Number of loci with a minimal length of 500bp present in all 28 diploid taxa, in at least 27 taxa (without missing outgroup), in at least 26 taxa (without missing outgroup), and in at least 25 taxa (without missing outgroup), respectively. Any regions within loci with coverage below two were removed.

Phylogenetics based on concatenation method

The concatenated matrix consists of 587 loci with a length of 426,709bp. For easier description and comparison, we divide the phylogenetic tree into five distinct clades. Clade I consists of section *Asperae* (subgenus *Aspera*) species and subgenus *Nipponobetula* whereas clade II comprises section *Lentae* (subgenus *Aspera*) and *B. michauxii*, a species of section *Apterocaryon* (subgenus *Betula*). Two species of subgenus *Acuminata* (*B. luminifera* and *B. hainanensis*) forms clade III which is sister to clade IV and clade V with the former including *B. humilis* (section *Apterocaryon*), *B. maximowicziana* (subgenus *Acuminata*) and three species of section *Costatae* (subgenus *Betula*) and the latter mainly including species of section *Betula* (subgenus *Betula*). *Betula nigra*, a species of section *Dahuricae* (subgenus *Betula*) is basal to clades IV and V (Fig. 5.4).

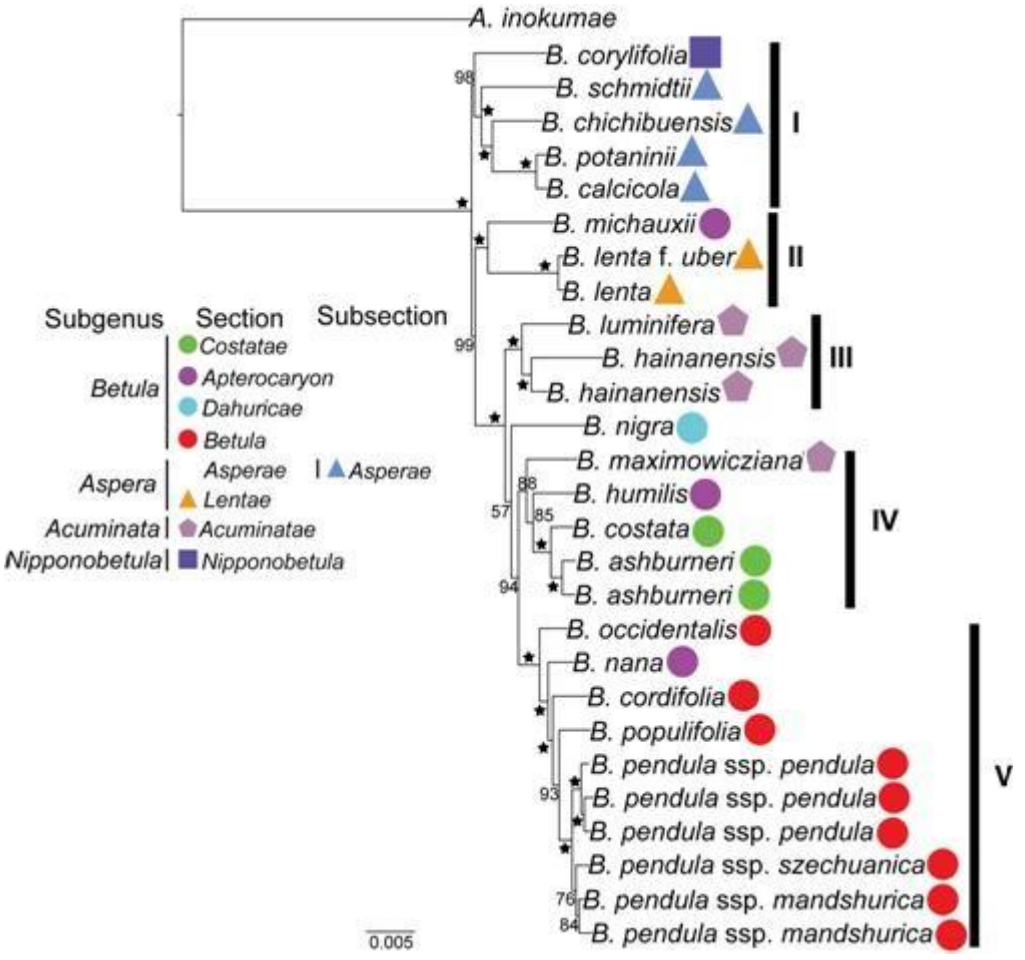


Figure 5.4 Maximum Likelihood analysis of diploid *Betula* species based on a concatenated matrix of 587 loci. Species are classified according to Ashburner & McAllister (2013). Marked with a star indicates a support value of 100%.

Phylogenetic tree based on STAR

Species tree inferred using STAR gives a highly-resolved phylogenetic tree which shares many similarities with the tree shown in Fig. 5.4. The species relationships within clade I, II, III and V are identical in both cases and nearly identical in clade IV except *B. maximowicziana* (Fig. 5.5). *Betula maximowicziana* is clustered within clade IV in the phylogentic tree shown in Fig. 5.4 whereas sister to clade IV and clade V (Fig. 5.5). In addition, clade I is sister to clade II whereas clade III is sister to clades IV, V and *B. maximowicziana* (Fig. 5.5). *Betula nigra* is basal to clades III, IV and V (Fig. 5.5).

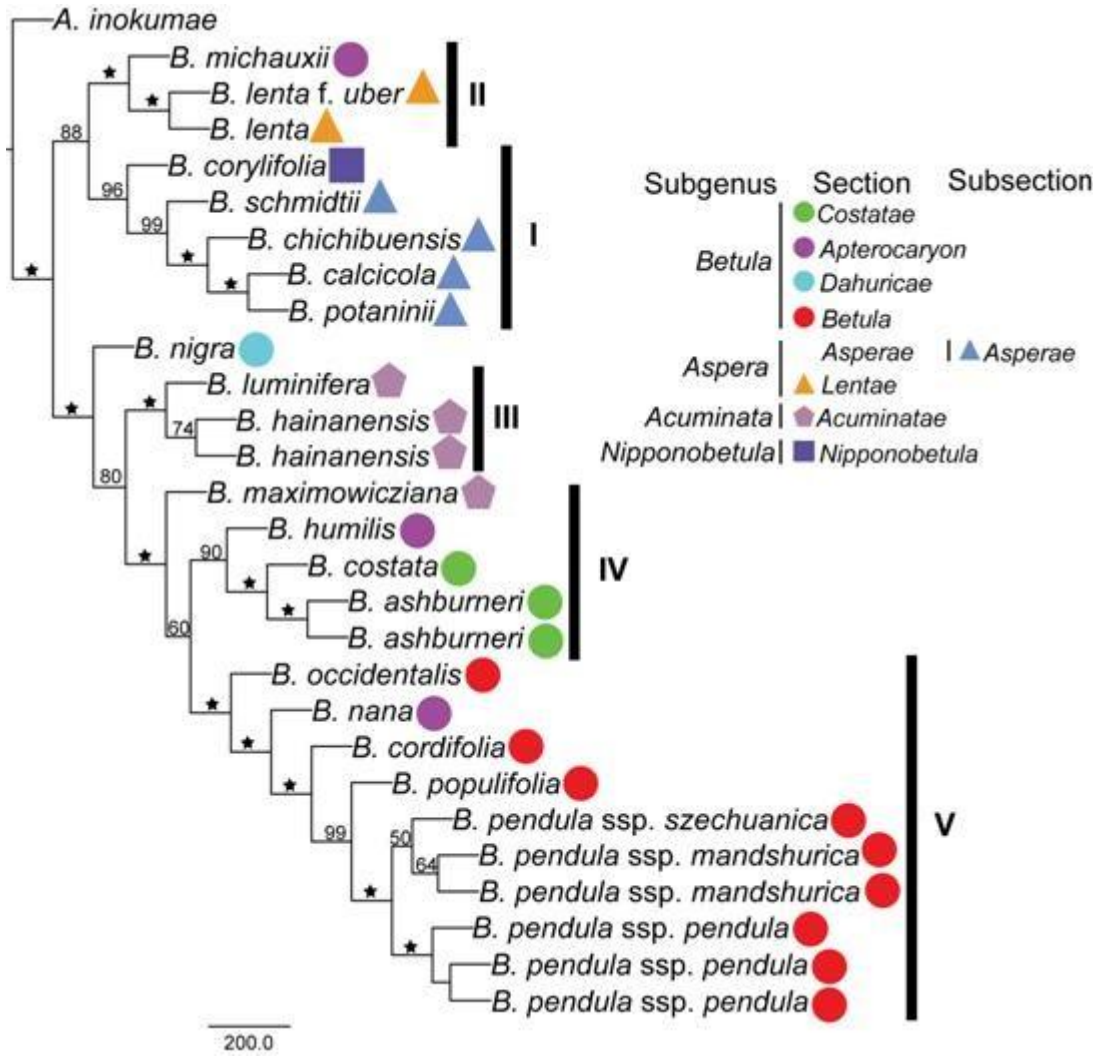


Figure 5.5 Species tree estimation using average ranks of coalescence (STAR) of diploid *Betula* species based on 587 loci. Species are classified according to Ashburner & McAllister (2013). Marked with a star indicates a support value of 100%.

Phylogenetic tree based on MP-EST

The species tree based on MP-EST supports the five clades (Fig. 5.6), within which the species relationships are identical compared with the tree constructed from the concatenated matrix (Fig. 5.4).

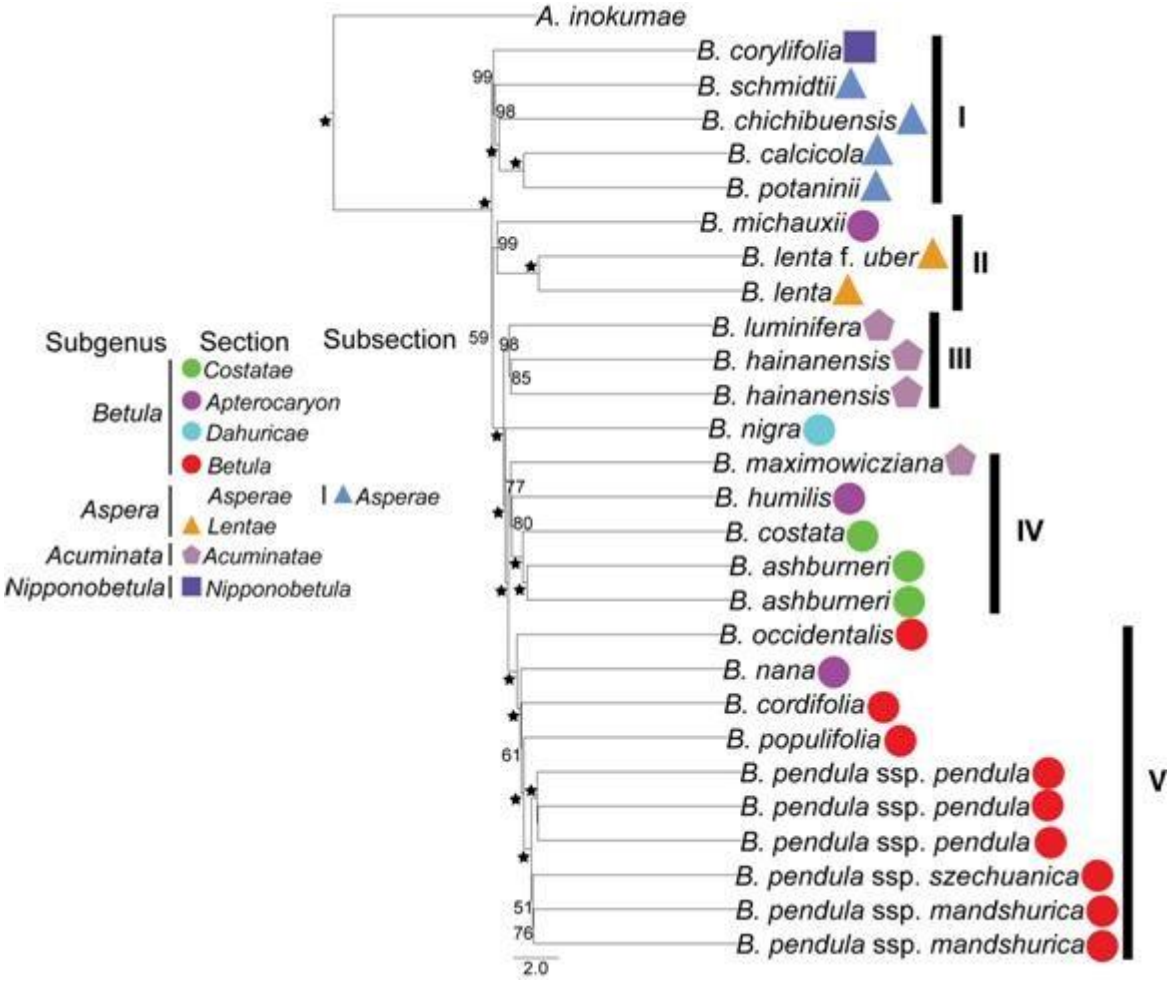


Figure 5.6 Maximum pseudo-likelihood estimation of species trees (MP-EST) of diploid *Betula* species based on 587 loci. Species are classified according to Ashburner & McAllister (2013). Marked with a star indicates a support value of 100%.

Phylogenetics based on binary data

The total number of binary characters for each taxon is 82,578. Phylogenetic analysis based on binary presence/absence data reveals four distinct clades, which are similar to the corresponding clades revealed by DNA characters but with a few exceptions. The phylogenetic position of *B. corylifolia* is uncertain based on binary data; it clusters with clade II with very weak support (below 50%). *Betula humilis* is nested within clade V forming a sister to *B. populifolia* whereas within which *B. nana* and *B.*

cordifolia form a sister. Clade III, which represented by *B. luminifera* and the two *B. hainanensis* individuals is not recovered based on binary data (Fig. 5.7). The phylogenetic position of *B. nigra* is identical as that of the STAR tree shown in Fig. 5.5.

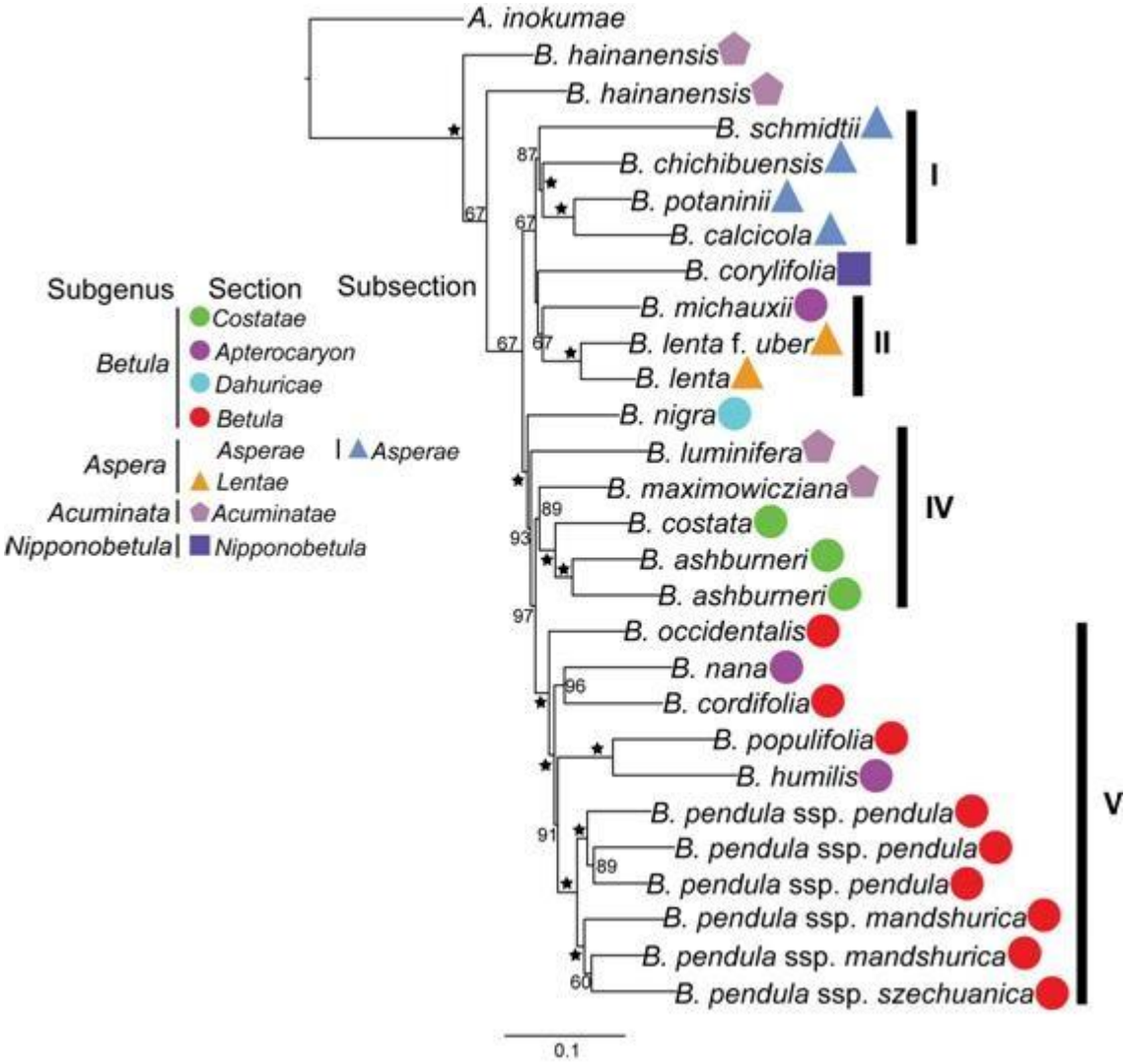


Figure 5.7 Maximum Likelihood analysis of diploid *Betula* species based on the binary presence/absence of RAD loci. Species are classified according to Ashburner & McAllister (2013). Marked with a star indicates a support value of 100%.

Discussion

Utility of RADSeq for phylogenomic analysis

This work presented here provides valuable genetic resources for *Betula*, and to a lesser extent, *Alnus* and *Corylus*. Here, we show that phylogenomic analysis based on RAD loci is feasible and facilitated by an available reference genome. Despite several disadvantages of RADSeq, such as uneven coverage among loci, RAD tags dropout and low number of RAD loci shared among distantly-related species, it is an effective method for study of genome-wide variation among species.

In this study, for most *Betula* species, the number of loci assembled with a minimal length of 500bp is over 15,000. This number is about 10,000 and 6,000 for *Alnus. inokumae* and *Corylus avellana*, respectively. Such a reduction in the number of loci for *Alnus* and *Corylus* compared with that of *Betula* species is due to their mapping to *B. platyphylla* genome but would increase significantly if the reads were assembled *de novo* or mapped to more closely-related genomes. In the present study, we just choose one *Alnus* species as the outgroup for phylogenetic analysis to maximise the number of usable loci. Seventy-three loci with a minimal length of 500bp are present in the 28 *Betula* and *Alnus* species, and this number increases significantly if missing taxa were allowed per locus. For example, this number would increase to 587 when no more than three missing *Betula* taxa are allowed per locus and this number increases dramatically if the proportion of missing data increases. In the dataset we use for phylogenomic study, the proportion of missing data is merely ~7.8%, far below that of other studies. Several previous studies have shown that missing data do not have a negative impact on phylogenetic analysis (Hovmöller *et al.*, 2013; Jiang *et al.*, 2014) especially when the total number of characters used increases (Wiens, 1998; Wiens, 2003). In contrast, excluding missing data may have a consequence on phylogenetic analysis as regions with missing data are probably more variable (Huang & Knowles, 2014). Species tree inference methods, such as STAR, MP-EST and NJst allows missing taxa. Given the fact that 587 loci with a minimal length of 500bp can be obtained allowing for three missing *Betula* taxa per locus, the number of loci derived by RADSeq is sufficient for phylogenomic analysis. Methods, such as NJst, allow not only for missing taxa but also for missing the outgroup in some loci (Liu & Yu, 2011). In this case, the number of useble loci is up to a few thousands in my dataset depending on the level of missing

data. Further research will be carried out to design the best strategy to maximise the futher use of the data presented in this chapter.

In this study, we use methods STAR and MP-EST to infer the species tree for the following reasons. First, these methods are based on summary statistics, which take little computational time and can be applied on a phylogenomic scale. In contrast, parameter-rich methods take up more computional resources and thus are not easily applicable at the genomic scale. In addition, the perfomance of parameter-rich methods can be compromised if loci have limited variations as this can cause failure in convergence (Knowles, 2009). Second, STAR takes account of deep coalescence of loci whereas MP-EST is robust to a small amount of gene flow or HGT (Liu *et al.*, 2009a; Liu *et al.*, 2010). Deep coalescence of loci and strong gene flow can be common for *Betula* because some species of this genus have large population size and long generation time, and hybridisation occurs. Moreover, simulation studies have shown that the two methods can reliably and consistently estimate a species tree when the number of loci tends toward infinity in theory (Liu *et al.*, 2009a; Liu *et al.*, 2010). In addition, we conducted phylogenetic analysis based on a matrix of binary presence/absence data: this gives a somewhat different result to those of STAR, MP-EST or the concatenation method, it still yields some strong phylogenetic signal. The main disadvantage of the presence/absence method is a high level of phylogenetic noise due to mutations of restriction cutting sites.

Phylogenetic relationships within **Betula**

Subgenus Aspera

Subgenus *Aspera* consists of section *Asperae* and section *Lentae* with the former including *B. calcicola*, *B. chichibuensis*, *B. potaninii* and *B. schmidtii* and with the latter including *B. lenta* and *B. lenta* f. *uber* based on STAR, MP-EST and the concatenation method. The two sections form a monophyletic group on their own based on all analyses, consistent with previous studies based on ITS sequences (Li *et al.*, 2005) and morphological characters (Ashburner & McAllister, 2013).

Subgenus Acuminata

Subgenus *Acuminata* consists of three diploid species: *B. hainanensis*, *B. luminifera* and *B. maximowicziana* (Chapter 4). The newly described species *B. hainanensis* forms a sister to *B. luminifera* based on STAR, MP-EST and concatenation methods whereas *B. maximowicziana* is more closely-related to species of section *Costatae*

(subgenus *Betula*). Similar placements have been suggested by phylogenetic analysis based on ITS sequences (Chapter 4). In addition, the autumn fruiting and much thicker male catkins of *B. maximowicziana* are distinct from other *Acuminata* species which may indicate a distant relationship. Of particular interest, *B. hainanensis* is restricted to Hainan island in China whereas *B. luminifera* is widespread across central and south China. A possible scenario is that some *B. luminifera* individuals colonised Hainan island giving rising to *B. hainanensis* later, or altenatively, some *B. hainanensis* individuals colonised mainland China, giving rise to *B. luminifera* as a result of adaption to temperate environment. Further research is needed to infer the evolutionary history of the two species.

Subgenus Betula

Subgenus *Betula* is divided into four sections: *Apterocaryon*, *Betula*, *Costatae* and *Dahuricae*. Most species of subgenus *Betula* form clade IV and clade V based on all analyses. STAR, MP-EST and concatenation method support the clustering of species of section *Apterocaryon*, represented by *B. humilis*, *B. nana* and *B. michauxii* into clade II, IV and V, respectively (Figs. 5.3-5.5). However, phylogenetic analysis of binary data indicates the clustering of *B. humilis* into clade V (Fig. 5.6). Classification of section *Apterocaryon* is simply based on their dwarfism, which is not a reliable character, as dwarfism can be caused by local adaption. Interestingly, *B. michauxii* is closely-related to *B. lenta*/*B. lenta* f. *uber*, which is consistent with a previous study based on *NIA* (Li *et al.*, 2007). Morphologically, *B. michauxii* is almost identical to *B. nana* but quite distantly-related to each other. *Betula costata* and *B. ashburneri*, two species of section *Costatae* reveals a close relationship, which is consistent with the recent monograph of *Betula* and with the result based on ITS. *Betula nigra* is an outlier to subgenus *Betula*, being placed as sister to most of species of subgenus *Betula* based on concatenation and MP-EST methods whereas it is basal to subgenus *Acuminata* and subgenus *Betula* based on STAR.

All species of section *Betula* form clade V, indicating their common ancestry. Interestingly, *B. pendula* are divided into two subclades: one including three *B. pendula* ssp. *pendula* native to Europe and another including *B. pendula* ssp. *szechuanica* from SW China and two *B. pendula* ssp. *mandshurica* from Japan and N. America. It has been suggested that *B. pendula* originates from eastern Asia, with one lineage dispersing into N. America whereas another dispersing into Europe. Further reseach is needed to compare the patterns of genetic diversity at the global scale.

Subgenus Nipponobetula

Subgenus *Nipponobetula* comprises the single species, *B. corylifolia*. It has some distinct morphological characters and thus been suggested as a subgenus (Skvortsov, 2002; Ashburner & McAllister, 2013). However, the results presented here show that *B. corylifolia* forms a strongly supported monophyletic group with species of section *Asperae* (subgenus *Aspera*). Similar results have been found based on ITS (Fig. 4.1), although weakly supported. Phylogenetic analysis of binary presence/absence data strongly supports the clustering of *B. corylifolia* within subgenus *Aspera* but does not resolve its position (Fig. 5.6). Despite some differences in morphological characters between *B. corylifolia* and species of section *Asperae*, its characters such as lack of seed wings and a single seed per bracts are similar. Hence, these characters may reflect the evolutionary relationships of these species and further research is needed to investigate the evolution of such morphological characters.

A new classification of* Betula *based on phylogenomic analysis

We propose a new classification of *Betula* based on these phylogenomic analyses. We use most diploid species other than polyploid species to avoid phylogenetic uncertainty due to reticulate evolution. In addition, we rely less than the recent monograph of *Betula* on morphological characters in classification, as they may be plastic or convergent.

Section *Apterocaryon* (*B. humilis*, *B. mchauxii* and *B. nana*), which is characterised by their dwarfism should be split and incorporated into other subgenera. *Betula michauxii* should belong to section *Asperae* (subgenus *Aspera*) whereas *B. humilis* and *B. nana* belong to subgenus *Betula*, with *B. humilis* being section *Costatae* and *B. nana* being section *Betula*. Subgenus *Nipponobetula*, represented by *B. corylifolia* should be incorporated into section *Asperae*. *Betula maximowicziana* should belong to subgenus *Betula* but its phylogenetic position is uncertain being sister to section *Costatae* based on MP-EST, concatenation method and binary presence/absence analyse whereas being sister to sections *Costatae* and *Betula* based on STAR. So, *B. maximowicziana* can be a new section within subgenus *Betula*. *Betula nigra*, placed within section *Dahuricae* (subgenus *Betula*) can be treated as a new subgenus given its phylogenetic position compared with other species of subgenus *Betula*.

Hence, we propose four subgenera and seven sections: *Acuminata* (section *Acuminatae* [*B. luminifera* and *B. hainanensis*]), *Aspera* (sections *Asperae* [*B. calcicola*, *B.*

chichibuensis, *B. corylifolia*, *B. potaninii* and *B. schmidtii*] and *Lentae* [*B. lenta*, *B. lenta* f. *uber* and *B. michauxii*]), *Betula* (sections *Betula* [*B. cordifolia*, *B. nana*, *B. occidentalis*, *B. pendula* ssp. and *B. populifolia*], *Costatae* [*B. costata* and *B. ashburneri*], and *Maximowiczianae* [*B. maximowicziana*]) and *Dahuricata* (section *Dahuricae* [*B. nigra*]). Subgenus *Acuminata* corresponds to section *Acuminatae* of Skvortsov (2002) and subgenus *Acuminata* of Ashburner and McAllister (2003) but excluding *B. maximowicziana* as it is closely-related to subgenus *Betula* based on all analyse. Subgenera *Aspera* and *Betula* are similar to these of Skvortsov (2002) and Ashburner and McAllister (2003). Within subgenus *Betula*, we removed previously proposed section *Apterocaryon* as its species are clustered into distinct clades in all phylogenomic analyses. In addition, we propose a new section *Maximowiczianae*, represented by *B. maximowicziana*. Section *Betula* and section *Costatae* largely agree with Skvortsov (2002) and Ashburner and McAllister (2003) but including *B. nana* and *B. humilis*, respectively. We propose section *Dahuricae* of Skvortsov (2002) and Ashburner and McAllister (2003) as subgenus *Dahuricata* as its representative species *B. nigra* is distantly-related to other subgenus *Betula* species based on all analyse.

Phylogenomic analyses do not support the division of *Betula* into two subgenera of Regel (1865): *Alnaster* and *Eubetula*, nor the division into two sections *Betulaster* and *Eubetula* of Winkler's classification (Winkler, 1904) as *Alnaster* or *Betulaster* represented by *B. luminifera*, *B. hainanensis* and *B. maximowicziana* are not monophyletic and the remaining species are neither. Our result does not support de Jong's classification of *Betula* into five subgenera as these all failed to be monophyletic. Skvortsov's classification of *Betula* into three subgenera and eight sections makes more sense. According to phylogenomic analysis based on STAR, subgenera *Asperae* (sections *Asperae*, *Lentae*, *Chinenses*) and *Betula* (sections *Acuminatae*, *Apterocaryon*, *Betula*, *Costatae* and *Dahuricae*) are monophyletic.

Chapter 6 Conclusions

General overview

The thesis provides new insights on three major topics: the introgression patterns among the three hybridising species: *B. nana*, *B. pubescens* and *B. pendula*, the phylogeny and genome size evolution of *Betula* based on ITS sequences and molecular phylogeny of diploid *Betula* species based on RAD loci. *Betula* is an emerging model for phylogenetic analysis, phylogeography and the evolution of polyploid species. This genus has the following attributes: (1) classification of species is difficult; (2) gene flow is common for species within subgenera or even among different subgenera; (3) polyploid species account for a majority of species, with the highest ploidy level reaching dodecaploidy and (4) monoploid genome size (1Cx-value) of *Betula* is small, ranging from ~0.38 pg to 0.60 pg (Chapter 4). The newly available genetic resources, such as the whole genome sequences of *B. nana* and RAD markers for nearly all *Betula* species, will make it an ideal model for future research.

The contribution of this thesis

Past climatic oscillations have had a huge impact on the current distribution of species, resulting in multiple refugia such as the peninsulas of Iberia, Italy, the Alps and the Balkans in Europe (Hewitt, 1999; Tribsch & Schonswetter, 2003), south-eastern North America (Soltis *et al.*, 2006), the Arctic and Beringia (Weider & Hobaek, 2000; Hewitt, 2004), southern Australia (Byrne, 2008), southeastern Asia (Gathorne-Hardy *et al.*, 2002; Meijaard, 2003) and Himalaya regions (Qiu *et al.*, 2011). Most species retreated into refugia during glaciation whereas they expanded their ranges during inter-glaciation periods as the climate warmed up.

The expansions and contractions of species ranges responding to these climatic changes have a great impact on shaping the genetic diversity and evolution of species (Hewitt, 1996). Recolonisation of habitats after glaciation often results in secondary contacts of different species, which may have been adapted to different environments during isolation. New species or combinations may form combining these different adaptations, which may be more suitable to novel environments. For example, it has been observed that higher frequencies of recent allopolyploid species occur in northern

parts, especially in arctic regions, possibly due in part to interspecific hybridisation of species or lineages from different refugia (Stebbins, 1984; Stebbins, 1985). So it is important to study speciation, phylogeography and parental origins of polyploid species in the context of past climate change. *Betula* provides excellent models to study these aspects in terms of past climate change as its species are widespread and have been shaped by past glaciation and polyploid species are common (Ashburner & McAllister, 2013).

In recent decades, with the advance of molecular tools, species demography and the postglacial history of species have come under extensive study (Petit *et al.*, 1997; Schmitt & Seitz, 2001; Petit *et al.*, 2003). For some species, refugial populations tend to have higher levels of genetic diversity compared with populations in colonised areas leaving a gradient of genetic diversity reflecting past colonisation routes (Petit *et al.*, 1997). However, sometimes the genetic diversity in colonised habitats is higher than that of the refugia populations, likely caused by the admixture of divergent lineages from separate refugia, as found in a study of European trees and shrubs (Petit *et al.*, 2003). During range expansion, closely-related species can hybridise when they come into contact. As a consequence of hybridisation, one species may eliminate another one via pollen swamping (Prentis *et al.*, 2007) or genetic assimilation (Levin *et al.*, 1996).

In Chapter 2, multiple lines of evidence show that hybridisation with *B. pubescens* partially explains the restricted distribution of *B. nana* in addition to range reduction due to climate change. These evidences include a cline of introgression from *B. nana* into *B. pubescens*, ecological niche modelling (ENM) and historical pollen/macrofossil records. For a locally endangered species, introgression of genetic material can help to understand the past species dynamics (Buggs, 2007; Currat *et al.*, 2008). We rule out the possibility of ancestral polymorphism in causing admixture between *B. nana* and *B. pubescens* as the north-to-south gradient of admixture from *B. nana* into *B. pubescens* is more likely caused by introgression when *B. pubescens* invades the range of *B. nana*. Pollen/macrofossil records reveal that *B. nana* was once widespread in Britian and its hybridisation with *B. pubescens* has occurred since the Holocene. So the picture is that after the last glaciation, *B. pubescens* colonised Britian following *B. nana* as the climate warmed up. During this period, hybridisation took place between the two species, causing the near-extinction of *B. nana* in England. A similar cline of introgression from *B. nana* into *B. pubescens* has also been detected at the European scale, indicating postglacial expansion of *B. pubescens* (Eidesen *et al.*, 2015). In

addition, we identified that although *B. nana* rarely hybridises with *B. pendula* due to geographical isolation, plantation of *B. pendula* near *B. nana* could cause a potential threat as hybrids have been detected between the two (Chapter 2). We also showed little genetic structure in *B. pubescens* and *B. pendula* across Britain, but distinct genetic structure in *B. nana*. This is plausible because *B. nana* is a shrub and its habitat in Britain has been heavily fragmented. Hence, gene flow between populations of *B. nana* is limited and genetic drift may play an important role in creating a distinct genetic structure. For *B. pendula* and *B. pubescens*, the lack of distinct genetic structures is possibly due to their long-distance dispersal and more recent colonisation. Similar results have been observed for *B. pubescens* and *B. pendula* even at the European scale (Maliouchenko *et al.*, 2007; Thórsson *et al.*, 2010).

Betula pubescens and *B. pendula* are morphologically similar. Atkinson devised a formula, namely, the Atkinson discriminant function (ADF) based on three leaf characters (Atkinson & Codling, 1986) to distinguish them. It was shown that *B. pubescens* tends to have an ADF score below zero, whereas *B. pendula* has one above zero. In Chapter 3, based on microsatellite analysis and ADF scores of 944 samples, we propose that the ADF score of -2 is a better boundary below which an individual is more likely to be *B. pubescens*. We suggest that the leaf variation within *B. pubescens* or *B. pendula* is possibly due to environmental factors other than gene flow.

In Chapter 4, I studied the phylogeny, genome size evolution and phylogeography of *Betula*. Compared with previous studies on the phylogenetics of *Betula* (Li *et al.*, 2005; Li *et al.*, 2007; Schenk *et al.*, 2008), I sampled nearly all described species and these species have been identified by the monograph author Hugh McAllister. This provides a baseline so that our discussions comparing morphological classification and phylogeny are meaningful. Moreover, the genome size of nearly all described species has been estimated, of which most are estimated for the first time, providing valuable knowledge for future research. The ITS phylogeny roughly corresponds with the classification given in a recent monograph of *Betula* (Ashburner & McAllister, 2013), although the proposed subgenera are not monophyletic and the relationships among subgenera and sections remain unresolved. However, the phylogenetic postions of a few species merit further investigation, such as *B. bomiensis*, *B. grossa*, *B. nigra*, *B. maximowicziana* and *B. michauxii*. We think that the unexpected phylogenetic positions of *B. bomiensis* and *B. grossa* are due to alloploy ploidy; and that of *B. michauxii* is attributed to morphological convergence in dwarfism. Our analysis

detected 24 taxa from botanical gardens or from GenBank with unexpected phylogenetic signals. This is likely caused by species misidentification or introgressive hybridisation. Also, we found that polyploid species with very high ploidy level (octoploidy and above) have very narrow ranges whereas some polyploid species, such as *B. pubescens* and *B. utilis* are widespread and invasive. It is possible that the restricted ranges of high ploidy level species are due to nutrition limitation, low dispersal ability or recent origins. An alternative explanation is that these polyploid species are largely autopolyploid species which may have lower adaptation ability even compared with their progenitor species (Stebbins, 1985).

Chapter 5 shows the power of phylogenomics in inferring the species relationships and opens up the possibility of future investigation of the topology of different gene trees in relation to function or evolutionary events. Although RADSeq has been commonly used for SNP discovery (Baird *et al.*, 2008), its utility for phylogenetics is in its infancy. Only in a few studies has RADSeq been adopted to infer shallow species relationships, such as in *Quercus* (Hipp *et al.*, 2014) and *Pedicularis* (Eaton & Ree, 2013). As far as I am aware, the study reported in this thesis is the first to demonstrate the use of RADSeq at the genus level with divergence between *Betula* and *Alnus* estimated to have occurred 60 Myr ago (Grimm & Renner, 2013). In addition, the advantage of our approach here over previous similar studies is three-fold. (1) We use MiSeq to generate long paired-end reads. Compared with data produced by HiSeq elsewhere (Cariou *et al.*, 2013; Hipp *et al.*, 2014; Pante *et al.*, 2015), the reads are much longer which is preferable for phylogenetic analyses due to the higher level of phylogenetic signal. In addition, longer reads may represent part of genes which are easier to functionally-annotate by blasting against known databases. (2) We used a reference genome for mapping. The advantage of mapping to a reference genome over de novo assembling is not only to separate homologous loci from paralogous loci but to anchor reads with adjacent restriction sites together. Such loci with a length of 500bp or greater are particularly important to infer gene trees and species trees using different methods. (3) Most phylogenetic studies using RAD thus far have relied upon short reads and concatenation of loci (a super-matrix approach): here I use a long-read approach that enables us to produce phylogenetic trees for each locus (a coalescence-based approach). The fact that I have different RAD loci with distinct tree topologies provides an opportunity for future study to compare different gene trees and to explore the function of such genes. I obtained RAD tags for nearly all described *Betula* species. By

mapping to the *B. platyphylla* genome, tens and thousands of loci with a minimal length of 500bp have been created for each species. I selected 587 loci to infer the phylogenetic relationships of diploid *Betula* species using various methods. The result shows that the recent classification of *Betula* proposed by Ashbuner and McAllister needs some minor revision. Based on it, I proposed a new classification for *Betula* which would be important for further research of this genus. In addition, RAD loci generated in this study provide a basis for inferring the parental of polyploid species and the evolution of genes for different lineages which can be facilitated by a reliably estimated phylogenetic tree using diploid species.

New questions and future research

Adaptive introgression

Betula provides a model system to investigate adaptive introgression as hybridisation and introgression occur extensively between species. Some species are extremely cold tolerant, such as the treeline species *B. nana*. In Europe, the rapid expansion of *B. pubescens* has been ascribed to adaptive introgression from *B. nana* (Eidesen *et al.*, 2015). However, the underling mechanism remains unclear. NGS technologies, such as RADSeq provide a possibility to detect alleles which are under strong natural selection. With the whole genome sequence of *Betula* species, it will be easier to identify the biological function of alleles under selection.

Conservation of endangered Betula species

Betula includes some critically endangered species, such as *B. calcicola*. For effective conservation, it is important to know the underling factors that are responsible for their endangeredness, such as habitat disturbance, climate change or hybridisation with closely-related species. Hybridisation and subsequent gene flow can cause a rare species to go extinct (Rhymer & Simberloff, 1996; Wolf *et al.*, 2001). This is likely to occur for some *Betula* species. Hence, it is necessary to assess to what extent hybridisation endangers a rare species. The methods used in this thesis are applicable to such studies. For example, ENM can be used to assess the ranges of the potentially hybridising species before the LGM, at the current time or in the future. Microsatellite markers are powerful in detecting the introgression patterns of the hybridising species across their distributions (Chapter 2). Flow cytometry is useful to confirm the hybrids of the hybridising species which are of differing ploidy level (Chapter 4).

Parental origins of polyploid species

Despite the importance of polyploidy in plants, it is a challenge to disentangle the origins of polyploid species. The lack of single- and low-copy nuclear genes can hamper the power in interpreting reticulate speciation events (Brysting *et al.*, 2011); moreover, shared standing variation, introgressive hybridisation and the occurrence of paralogs can result in incongruence among different loci (Linder & Rieseberg, 2004). The commonly used nrITS region has limited power if concerted evolution towards one parent occurs (Álvarez & Wendel, 2003). Parental origins of polyploid species remain unknown for *Betula* species due to lack of such studies. In Chapter 5, I developed tens and thousands of loci (Table 5.1) that may be used to determine the parental origins of polyploid species of this genus. With these genetic resources available, we are confident to obtain the best inference of the parental lineages of polyploid species. Based on this, a series of questions can be addressed. For example, in what way are polyploid species differently adapted from their parents? What are the impacts of past climatic oscillations on the current distribution of polyploid species and their parents? Does the degree of parental divergence influence the formation of polyploid species?

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman JW, *et al.* 2013. Hybridization and speciation. *Journal of Evolutionary Biology* **26:** 229–246.

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29:** 417–434.

Anamthawat-Jónsson K, Tómasson T. 1990. Cytogenetics of hybrid introgression in Icelandic birch. *Hereditas* **112:** 65–70.

Anamthawat-Jónsson K, Tómasson T. 1999. High frequency of triploid birch hybrid by *Betula nana* seed parent. *Hereditas* **130:** 191–193.

Anamthawat-Jónsson K, Thórsson *Æ*, Temsch EM, Greilhuber J. 2010. Icelandic birch polyploids-the case of perfect fit in genome size. *Journal of Botany* **347254**.

Anamthawat-Jónsson K, Thórsson AT. 2003. Natural hybridisation in birch: triploid hybrids between *Betula nana* and *B. pubescens*. *Plant Cell Tissue and Organ Culture* **75:** 99–107.

Anderson E 1949. *Introgressive hybridisation.* New York: John Wiley.

Anderson E. 1953. Introgressive hybridization. *Biological Reviews* **28:** 280–307.

Andrews S. 2014. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arbogast BS, Kenagy GJ. 2001. Comparative phylogeography as an integrative approach to historical biogeography. *Journal of Biogeography* **28:** 819–825.

Arnold ML, Bulger MR, Burke JM, Hempel AL, Williams JH. 1999. Natural hybridization: how low can you go and still be important? *Ecology* **80:** 371–381.

Ashburner K, McAllister HA 2013. *The genus Betula: a taxonomic revision of birches.* London: Kew Press.

Aston D. 1984. *Betula nana* L., a note on its status in the United Kingdom. *Proceedings of the Royal Society of Edinburgh Section B-Biological Sciences* **85:** 43–47.

Atkinson MD. 1992. *Betula pendula* Roth (*B. verrucosa* Ehrh) and *B. pubescens* Ehrh. *Journal of Ecology* **80:** 837–870.

Atkinson MD, Codling AN. 1986. A reliable method for distinguishing between *Betula pendula* and *B. pubescens*. *Watsonia* **7:** 5–76.

Avise JC 2000. *Phylogeography: the history and formation of species.* Cambridge, MA: Harvard University Press.

Bai CK, Alverson WS, Follansbee A, Waller DM. 2012. New reports of nuclear DNA content for 407 vascular plant taxa from the United States. *Annals of Botany* **110:** 1623–1629.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One* **3:** e3376.

Balao F, Casimiro-Soriguer R, Garcia-Castano JL, Terrab A, Talavera S. 2015. Big thistle eats the little thistle: does unidirectional introgressive hybridization endanger the conservation of *Onopordum hinojense*? *New Phytologist* **206:** 448–458.

Barchi L, Lanteri S, Portis E, Acquadro A, Valè G, Toppino L, Rotino GL. 2011. Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *Bmc Genomics* **12:** 304.

Barnes BV, Bruce PD, Sharik TL. 1974. Natural hybridization of yellow birch and white birch. *Forest Science* **20:** 215–221.

Barnes BV, Dancik BP. 1985. Characteristics and origin of a new birch species, *Betula murrayana*, from southeastern Michigan. *Canadian Journal of Botany* **63:** 223–226.

Barriball K, McNutt EJ, Gorchov DL, Rocha OJ. 2015. Inferring invasion patterns of *Lonicera maackii* (Rupr) Herder (Caprifoliaceae) from the genetic structure of 41 naturalized populations in a recently invaded area. *Biological Invasions* **17:** 2387–2402.

Barton NH, Hewitt GM. 1985. Analysis of hybrid zones. *Annual Review of Ecology and Systematics* **16:** 113–148.

Barton NH, Hewitt GM. 1989. Adaptation, speciation and hybrid zones. *Nature* **341:** 497–503.

Baum DA, Small RL, Wendel JF. 1998. Biogeography and floral evolution of Baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. *Systematics Biology* **47:** 181–207.

Bennett MD, Leitch IJ. 2010. Plant DNA C-values Database (release 5.0, December 2010). http://data.kew.org/cvalues/.

Bennett MD, Leitch IJ. 2011. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Annals of Botany* **107:** 467–590.

Bennett MD, Smith JB. 1991. Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **334:** 309–345.

Bennetzen JL, Ma JX, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Annals of Botany* **95:** 127–132.

Blackburn KB. 1952. The dating of a deposit containing an elk skeleton found at Neasham, near Darlington, County Durham. *New Phytologist* **51:** 364.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120.

Bolnick DI, Turelli M, Lopez-Fernandez H, Wainwright PC, Near TJ. 2008. Accelerated mitochondrial evolution and "Darwin's corollary": asymmetric viability of reciprocal F1 hybrids in centrarchid fishes. *Genetics* **178:** 1037–1048.

Boni MF, Posada D, Feldman MW. 2007. An exact non-parametric method for inferring mosaic structure in sequence triplets. *Genetics* **176:** 1035–1047.

Bouillé M, Bousquet J. 2005. Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): Implications for the long-term maintenance of genetic diversity in trees. *American Journal of Botany* **92:** 63–73.

Bousquet J, Strauss SH, Li P. 1992. Complete congruence between morphological and rbcL-based molecular phylogenies in birches and related Species (Betulaceae). *Molecular Biology and Evolution* **9:** 1076–1088.

Bradshaw RHW. 1981. Modern pollen-representation factors for woods in south-east England. *Journal of Ecology* **69:** 45–70.

Britch SC, Cain ML, Howard DJ. 2001. Spatio-temporal dynamics of the *Allonemobius fasciatus*–A. socius mosaic hybrid zone: a 14-year perspective. *Molecular Ecology* **10:** 627–638.

Brown IR, Aldawoody D. 1979. Observations on meiosis in three cytotypes of *Betula alba* L. *New Phytologist* **83:** 801–811.

Brown IR, Kennedy D, Williams DA. 1982. The occurrence of natural hybrids between *Betula pendula* Roth and *B. pubescens* Ehrh. *Watsonia* **14:** 133–145.

Brown IR, Tuley G. 1971. A study of a population of birches in Glen Gairn. *Botanical Journal of Scotland* **41:** 231–245.

Brown IR, Williams DA. 1984. Cytology of *Betula alba* L. complex. *Proceedings of the Royal Society of Edinburgh Section B-Biological Sciences* **85:** 49–64.

Bruvo R, Michiels NK, D'Souza TG, Schulenburg H. 2004. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology* **13:** 2101–2106.

Brysting AK, Mathiesen C, Marcussen T. 2011. Challenges in polyploid phylogenetic reconstruction: a case story from the arctic-alpine *Cerastium alpinum* complex. *Taxon* **60:** 333–347.

Buerkle CA, J. MR, Asmussen MA, Rieseberg LH. 2000. The likelihood of homoploid hybrid speciation. *Heredity* **84:** 441–451.

Buggs R, Chamala S, Wu W, Tate J, Schnable P, Soltis D, Soltis P, Barbazuk W. 2012a. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Current Biology* **22:** 248–252.

Buggs RJA. 2007. Empirical study of hybrid zone movement. *Heredity* **99:** 301–312.

Buggs RJA, Chamala S, Wu W, Tate JA, Schnable PS, Soltis DE, Soltis PS, Barbazuk WB. 2012b. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of Independent origin. *Current Biology* **22:** 248–252.

Buggs RJA, Doust AN, Tate JA, Koh J, Soltis K, Feltus FA, Paterson AH. 2009. Gene loss and silencing in *Tragopogon miscellus* (Asteraceae): comparison of natural and synthetic allotetraploids. *Heredity* **103:** 73–81.

Buggs RJA, Elliott NM, Zhang LJ, Koh J, Viccini LF, Soltis DE, Soltis PS. 2010. Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. *New Phytologist* **186:** 175–183.

Buggs RJA, Pannell JR. 2006. Rapid displacement of a monoecious plant lineage is due to pollen swamping by a dioecious relative. *Current Biology* **16:** 996–1000.

Burgess KS, Morgan M, Deverno L, Husband BC. 2005. Asymmetrical introgression between two *Morus* species (*M. alba*, *M. rubra*) that differ in abundance. *Molecular Ecology,* **14:** 3471–3483.

Byrne M. 2008. Evidence for multiple refugia at different time scales during Pleistocene climatic oscillations in southern Australia inferred from phylogeography. *Quaternary Science Reviews* **27:** 2576–2585.

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR. 1993. Incidence and origin of null alleles in the $(AC)_n$ microsatellite markers. *American Journal of Human Genetics* **52:** 922–927.

Campbell DR. 2004. Natural selection in *Ipomopsis* hybrid zones: implications for ecological speciation. *New Phytologist* **161:** 83–90.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972–1973.

Cariou M, Duret L, Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution* **3:** 846–852.

Caseldine C. 2001. Changes in *Betula* in the Holocene record from Iceland - a palaeoclimatic record or evidence for early Holocene hybridisation? *Review of Palaeobotany and Palynology* **117:** 139–152.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22:** 3124–3140.

Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V. 2005. Land plants and DNA barcodes: short-term and long-term

goals. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360:** 1889–1895.

Chen IC, Hill JK, Ohlemuller R, Roy DB, Thomas CD. 2011. Rapid range shifts of species associated with high levels of climate warming. *Science* **333:** 1024–1026.

Chen J, Kallman T, Gyllenstrand N, Lascoux M. 2010. New insights on the speciation history and nucleotide diversity of three boreal spruce species and a Tertiary relict. *Heredity* **104:** 3–14.

Chen ZD, Manchester SR, Sun HY. 1999. Phylogeny and evolution of the Betulaceae as inferred from DNA sequences, morphology, and paleobotany. *American Journal of Botany* **86:** 1168–1181.

Chester M, Gallagher JP, Symonds VV, da Silva AVC, Mavrodiev EV, Leitch AR, Soltis PS, *et al*. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proceedings of the National Academy of Sciences of the United States of America* **109:** 1176–1181.

Clark LV, Jasieniuk M. 2011. POLYSAT: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources* **11:** 562–566.

Clark LV, Stewart JR, Nishiwaki A, Toma Y, Kjeldsen JB, Jørgensen U, Zhao H, *et al*. 2015. Genetic structure of *Miscanthus sinensis* and *Miscanthus sacchariflorus* in Japan indicates a gradient of bidirectional but asymmetric introgression. *Journal of Experimental Botany* **66:** 4213–4225.

Cowan RS, Chase MW, Kress WJ, Savolainen V. 2006. 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon* **55:** 611–616.

Crawford RMM 2008. *Plants at the margin: ecological limits amd climate change.* Cambridge: Cambridge University Press.

Cruaud A, Gautier M, Galan M, Foucaud J, Saune L, Genson G, Dubois E, Nidelet S, Deuve T, Rasplus JY. 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution* **31:** 1272–1274.

Cruzan MB. 2005. Patterns of introgression across an expanding hybrid zone: analysing historical patterns of gene flow using nonequilibrium approaches. *New Phytologist* **167:** 267–278.

Cuenca J, Aleza P, Navarro L, Ollitrault P. 2013. Assignment of SNP allelic configuration in polyploids using competitive allele-specific PCR: application to citrus triploid progeny. *Annals of Botany* **111:** 731–742.

Currat M, Excoffier L. 2011. Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 15129–15134.

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. *Evolution* **62:** 1908–1920.

Czernicka M, Pławiak J, Muras P. 2014. Genetic diversity of F1 and F2 interspecific hybrids between dwarf birch (*Betula nana* L.) and Himalayan birch (*B. utilis* var. *jacquemontii* (Spach) Winkl. 'Doorenbos') using RAPD-PCR markers and ploidy analysis. *Acta Biochimica Polonica* **61:** 195–199.

Dakin EE, Avise JC. 2004. Microsatellite null alleles in parentage analysis. *Heredity* **93:** 504–509.

Dancik BP, Barnes BV. 1972. Natural variation and hybridization of yellow birch and bog birch in southeastern Michigan. *Silvae Genetica* **21:** 1–9.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9:** 772–772.

Davis MB, Shaw RG. 2001. Range shifts and adaptive responses to Quaternary climate change. *Science* **292:** 673–679.

Davy AJ, Gill JA. 1984. Variation due to environment and heredity in birch transplanted between heath and bog. *New Phytologist* **97:** 489–505.

Day PD, Berger M, Hill L, Fay MF, Leitch AR, Leitch IJ, Kelly LJ. 2014. Evolutionary relationships in the medicinally important genus *Fritillaria* L. (Liliaceae). *Molecular Phylogenetics and Evolution* **80:** 11–19.

De Jong PC. 1993. An introduction to *Betula*: its morphology, evolution, classification and distribution, with a survey of recent work. In: Hunt, D ed *Proceedings of the IDS Betula symposium, 2–4 October 1992. International Dendrology Society*. Richmond, UK.

de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. *Trends in Ecology & Evolution* **22:** 34–41.

de Queiroz K. 2007. Species concepts and species delimitation. *Systematic Biology* **56:** 879–886.

Degnan JH, Rosenberg NA. 2006. Discordance of species tree with their most likely gene trees. *Plos Genetics* **2:** 762–768.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24:** 332–340.

DeGroot WJ, Thomas PA, Wein RW. 1997. *Betula nana* L. and *Betula glandulosa* Michx. *Journal of Ecology* **85:** 241–264.

Dehond PE, Campbell CS. 1989. Multivariate analyses of hybridization between *Betula cordifolia* and *B. populifolia* (Betulaceae). *Canadian Journal of Botany* **67:** 2252–2260.

Doležel J, Bartos J. 2005. Plant DNA flow cytometry and estimation of nuclear genome size. *Annals of Botany* **95:** 99–110.

Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R. 1998. Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Annals of Botany* **82:** 17–26.

Doležel J, Greilhuber J, Suda J. 2007. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* **2:** 2233–2244.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* **284** 2124–2128.

Dowling TE, Demarais BD. 1993. Evolutionary significance of introgressive hybridization in cyprinid fishes. *Nature* **362:** 444–446.

Dowling TE, Secor CL. 1997. The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics* **28:** 593–619.

Dray S, Dufour AB. 2007. The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software* **22:** 1–20.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28:** 2239–2252.

Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematics Biology* **62:** 689 – 706.

Edwards C, Soltis DE, Soltis PS. 2008. Using patterns of genetic structure based on microsatellite loci to test hypotheses of current hybridization, ancient hybridization and incomplete lineage sorting in *Conradina* (Lamiaceae). *Molecular Ecology* **17:** 5157–5174.

Eidesen PB, Alsos IG, Brochmann C. 2015. Comparative analyses of plastid and AFLP data suggest different colonization history and asymmetric hybridisation between *Betula pubescens* and *B. nana*. *Molecular Ecology* **24:** 3993–4009.

Eidesen PB, Ehrich D, Bakkestuen V, Alsos IG, Gilg O, Taberlet P, Brochmann C. 2013. Genetic roadmap of the Arctic: plant dispersal highways, traffic barriers and capitals of diversity. *New Phytologist* **200:** 898–910.

Eldredge N, Cracraft J 1980. *Phylogenetic patterns and the evolutionary process.* New York: Columbia University Press.

Elith J, Leathwick JR. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics* **40:** 677–697.

Elkington TT. 1968. Introgressive hybridization between *Betula nana* L. and *B. pubescens* Ehrh. in North-West Iceland. *New Phytologist* **67:** 109–118.

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* **16:** 551–558.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5:** 435–445.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107:** 16196–16200.

Erdogan V, Mehlenbacher SA. 2000. Phylogenetic relationships of *Corylus* species (Betulaceae) based on nuclear ribosomal DNA ITS region and chloroplast matK gene sequences. *Systematic Botany* **25:** 727–737.

Estoup A, Jarne P, Cornuet JM. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11:** 1591–1604.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA. 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: Orgogonzo, V, Rockman, MV eds. *Molecular Methods for Evolutionary Genetics.* NY: Humana Press.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14:** 2611–2620.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceedings of the National Academy of Sciences of the United States of America* **106:** 5737.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17:** 368–376.

Felsenstein J. 1989. PHYLIP – Phylogeny inference package (version 3.6). *Cladistics* **5:** 164–166.

Fitzpatrick BM, Johnson JR, Kump DK, Smith JJ, Voss SR, Shaffer HB. 2010. Rapid spread of invasive genes into a threatened native species. *Proceedings of the National Academy of Sciences of the United States of America* **23:** 3606–3610.

Forest F, Bruneau A. 2000. Phylogenetic analysis, organization, and molecular evolution of the nontranscribed spacer of 5S ribosomal RNA genes in *Corylus* (Betulaceae). *International Journal of Plant Sciences* **161:** 793–806.

Forest F, Savolainen V, Chase MW, Lupia R, Bruneau A, Crane PR, Lavin M. 2005. Teasing apart molecular-versus fossil-based error estimates when dating

phylogenetic trees: a case study in the birch family (Betulaceae). *Systematic Botany* **30:** 118–133.

Franiel I, Więski K. 2005. Leaf features of silver birch (*Betula pendula* Roth). Variability within and between two populations (uncontaminated vs Pb-contaminated and Zn-contaminated site). *Trees* **19:** 81–88.

Fridley JD, Craddock A. 2015. Contrasting growth phenology of native and invasive forest shrubs mediated by genome size. *New Phytologist* **207:** 659–668.

Fuentes I, Stegemann S, Golczyk H, Karcher D, Bock R. 2014. Horizontal genome transfer as an asexual path to the formation of new species. *Nature* **511:** 232–235.

Furlow J. 1990. The genera of Betulaceae in the south eastern United States. *Journal of Arnold Arboretum* **71:** 1–67.

Furrer R, Nychka D, Sain S. 2011. Fields: Tools for spatial data. R package.

Gathorne-Hardy FJ, Syaukani, Davies RG, Eggleton F, Jones DT. 2002. Quaternary rainforest refugia in south-east Asia: using termites (Isoptera) as indicators. *Biological Journal of the Linnean Society* **75:** 453–466.

Gavin DG, Fitzpatrick MC, Gugger PF, Heath KD, Rodrıguez-Sanchez F, Dobrowski SZ, Arndt Hampe A, *et al.* 2014. Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist* **204:** 37–54.

Genner MJ, Turner GF. 2012. Ancient hybridization and phenotypic novelty within lake Malawi's cichlid fish radiation. *Molecular Biology and Evolution* **29:** 195–206.

Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16:** 573–582.

Gilbert KJ, Andrew RL, Bock DG, Franklin MT, Kane NC, Moore J, Moyers BT, *et al.* 2012. Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. *Molecular Ecology* **21:** 4925–4930.

Gill JA, Davy AJ. 1983. Variation and polyploidy within lowland populations of the *Betula pendula*/*Betula pubescens* complex. *New Phytologist* **94:** 433–451.

Gimingham CH. 1984. Ecological aspects of birch. *Proceedings of the Royal Society of Edinburgh Section B-Biological Sciences* **85B:** 65–72.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, *et al.* 2010. A draft sequence of the Neandertal genome. *Science* **328:** 710–722.

Gregory TR. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews* **76:** 65–101.

Greilhuber J. 2008. Cytochemistry and C-values: the less-well-known world for nuclear DNA amounts. *Annals of Botany* **101:** 791–804.

Greilhuber J, Dolezel J, Lysak MA, Bennett MD. 2005. The origin, evolution and proposed stabilization of the terms 'Genome Size' and 'C-Value' to describe nuclear DNA contents. *Annals of Botany* **95:** 91–98.

Grimaldi MC, Crouau-Roy B. 1997. Microsatellite allelic homoplasy due to variable flanking sequences. *Journal of Molecular Evolution* **44:** 336–340.

Grime JP, Mowforth MA. 1982. Variation in genome size – an ecological interpretation. *Nature* **299:** 151–153.

Grimm GW, Renner SS. 2013. Harvesting Betulaceae sequences from GenBank to generate a new chronogram for the family. *Botanical Journal of the Linnean Society* **172:** 465–477.

Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, Lepais O, Lepoittevin C, *et al*. 2011. Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11:** 591–611.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52:** 696–704.

Hörandl E. 2006. Paraphyletic versus monophyletic taxa - evolutionary versus cladisticclassifications. *Taxon* **55:** 564–570.

Hörandl E, Stuessy TF. 2010. Paraphyletic groups as natural units of biological classification. *Taxon* **59:** 1641–1653.

Haasl RJ, Payseur BA. 2011. Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity* **106:** 158–171.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN. 2006. DNA barcodes distinguish species of tropical *Lepidoptera*. *Proceedings of the National Academy of Sciences of the United States of America* **103:** 968–971.

Hall BG 2004. *Phylogenetic trees made easy: A how-to manual.* Sunderland (MA): Sinauer Associates.

Hammer Ø, Harper DAT, Ryan PD. 2001. PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4:** 1–9.

Harrison RG. 1990. Hybrid zones: windows on evolutionary process. *Oxford Surveys in Evolutionary Biology* **7:** 69–128.

Harrison RG, Rand DM. 1989. Mosaic hybrid zone and the nature of species boundaries. In: Otte, D, Endler, JA eds. *Speciation and its consequences.* Sinauer, Sunderland, M. A.

Hausdorf B. 2011. Progress toward a general species concept. *Evolution* **65:** 923–931.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. 2004a. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America* **101:** 14812–14817.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. 2004b. Identification of birds through DNA barcodes. *Plos Biology* **2:** 1657–1663.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405:** 907–913.

Hewitt GM. 1988. Hybrid zones - natural laboratories for evolutionary studies. *Trends in Ecology & Evolution* **3:** 158–167.

Hewitt GM. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58:** 247–276.

Hewitt GM. 1999. Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society* **68:** 87–112.

Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **359:** 183–195.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25:** 1965–1978.

Himes CMT, Gallardo MH, Kenagy GJ. 2008. Historical biogeography and post-glacial recolonization of South American temperate rain forest by the relictual marsupial *Dromiciops gliroides*. *Journal of Biogeography* **35:** 1415–1424.

Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, Manos PS. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. *Plos One* **9:** e93975.

Hoffman JI, Amos W. 2005. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology* **14:** 599–612.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *Plos Genetics* **6:** e1000862.

Holm SO. 1994. Reproductive patterns of *Betula pendula* and *B. pubescens* Coll along a regional altitudinal gradient in Northern Sweden. *Ecography* **17:** 60–72.

Hovmöller R, Knowles LL, Kubatko LS. 2013. Effects of missing data on species tree estimation under the coalescent. *Molecular Phylogenetics and Evolution* **69:** 1057–1062.

Howland DE, Oliver RR, Davy AJ. 1995. Morphological and molecular variation in natural populations of *Betula*. *New Phytologist* **130:** 117–124.

Huang H, Knowles LL. 2014. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Systematics Biology* **syu046**.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information *Molecular Ecology Resources* **9:** 1322–1332.

Huelsenbeck JP, Bull JJ, Cunningham CW. 1996. Combining data in phylogenetic analysis. *Trends in Ecology & Evolution* **11:** 152–158.

Hui W, Gel YR, Gastwirth JL. 2008. lawstat: an R package for law, public policy and biostatistics. *Journal of Statistical Software* **28:** 1–26.

Huntley B, Birks HJB 1983. *An atlas of past and present pollen maps for Europe: 0-13000 years ago.* Cambridge: Cambridge University Press.

Huxel GR. 1999. Rapid displacement of native species by invasive species: effects of hybridization. *Biological Conservation* **89:** 143–152.

Järvinen P, Palmé AE, Morales LO, Lännenpää M, Keinänen M, Sopanen T, Lascoux M. 2004. Phylogenetic relationships of *Betula* species (Betulaceae) based on nuclear ADH and chloroplast matK sequences. *American Journal of Botany* **91:** 1834–1845.

Jadwiszczak K, Banaszek A, Jabłońska E, Sozinov O. 2012. Chloroplast DNA variation of *Betula humilis* Schrk. in Poland and Belarus. *Tree Genetics & Genomes* **8:** 1017–1030.

Jakob SS, Ihlow A, Blattner FR. 2007. Combined ecological niche modelling and molecular phylogeography revealed the evolutionary history of Hordeum marinum (Poaceae) - niche differentiation, loss of genetic diversity, and speciation in Mediterranean Quaternary refugia. *Molecular Ecology* **16:** 1713–1727.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23:** 1801–1806.

Jaroszeski MJ, Radcliff G. 1999. Fundamentals of flow cytometry. *Molecular Biotechnology* **11:** 37–53.

Jiang W, Chen SY, Wang H, Li DZ, Wiens JJ. 2014. Should genes with missing data be excluded from phylogenetic analyses? *Molecular Phylogenetics and Evolution* **80:** 308–318.

Jiao YN, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, *et al*. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473:** 97–100.

Johnson PCD, Haydon DT. 2006. Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. *Genetics* **175:** 827–842.

Johnsson H. 1945. Interspecific hybridization within the genus *Betula*. *Hereditas* **31:** 163–176.

Jones AG, Ardren WR. 2003. Methods of parentage analysis in natural populations. *Molecular Ecology* **12:** 2511–2523.

Karlsdottir L, Hallsdottir M, Thórsson Æ, Anamthawat-Jónsson K. 2009. Evidence of hybridisation between *Betula pubescens* and *B. nana* in Iceland during the early Holocene. *Review of Palaeobotany and Palynology* **156:** 350–357.

Karlsson PS, Schleicher LF, Weih M. 2000. Seedling growth characteristics in three birches originating from different environments. *Ecoscience* **7:** 80–85.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33:** 511–518.

Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novák P, Neumann P, Lysak MA, *et al*. 2015. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytologist* **DOI: 10.1111/nph.13471**.

Kenworthy JB, Aston D, Bucknall SA. 1972. A study of hybrids between *Betula pubescens* Ehrh. and *Betula nana* L. from Sutherland - an integrated approach. *Transactions of the Botanical Society of Edinburgh* **41:** 517–539.

Kloda JM, Dean PDG, Maddren C, MacDonald DW, Mayes S. 2008. Using principle component analysis to compare genetic diversity across polyploidy levels within plant complexes: an example from British Restharrows (*Ononis spinosa* and *Ononis repens*). *Heredity* **100:** 253–260.

Knight CA, Molinari NA, Petrov DA. 2005. The large genome constraint hypothesis: evolution, ecology, and phenotype. *Annals of Botany* **95:** 177–190.

Knowles LL. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Systematics Biology* **58:** 463–467.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* **39:** 309–338.

Koropachinskii IY. 2013. Natural hybridization and taxonomy of birches in North Asia. *Contemporary Problems in Ecology* **6:** 350–369.

Kovacic S, Nikolic T. 2005. Relations between *Betula pendula* Roth.(Betulaceae) leaf morphology and environmental factors in five regions of Croatia. *Acta Bioogica Cracoviensia* **47:** 7–13.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. 2005. Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 8369–8374.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* **25:** 971–973.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematics Biology* **56:** 17–24.

Kulju KKM, Pekkinen M, Varvio S. 2004. Twenty-three microsatellite primer pairs for *Betula pendula* (Betulaceae). *Molecular Ecology Notes* **4:** 471–473.

Lavergne S, Muenke NJ, Molofsky J. 2010. Genome size reduction can trigger rapid phenotypic evolution in invasive plants. *Annals of Botany* **105:** 109–116.

Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid plants. *Science* **320:** 481–483.

Leitch AR, Leitch IJ. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist* **194:** 629–646.

Lepais O, Petit RJ, Guichoux E, Lavabre JE, Alberto F, Kremer A, Gerber S. 2009. Species relative abundance and direction of introgression in oaks. *Molecular Ecology* **18:** 2228–2242.

Levin DA, Francisco‐Ortega J, Jansen RK. 1996. Hybridization and the extinction of rare plant species. *Conservation Biology* **10:** 10‐16.

Lewis D, Crowe LK. 1958. Unilateral interspecific incompatibility in flowering plants. *Heredity* **12:** 233–256.

Lewontin RC, Birch LC. 1966. Hybridization as a source of variation for adaptation to new environments. *Evolution* **20:** 315–336.

Lexer C, Buerkle CA, Joseph JA, Heinze B, Fay MF. 2007. Admixture in European Populus hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Heredity* **98:** 74–84.

Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, Chen ZD, *et al*. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 19641–19646.

Li JH, Shoup S, Chen ZD. 2005. Phylogenetics of *Betula* (Betulaceae) inferred from sequences of nuclear ribosomal DNA. *Rhodora* **107:** 69–86.

Li JH, Shoup S, Chen ZD. 2007. Phylogenetic relationships of diploid species of *Betula* (Betulaceae) inferred from DNA sequences of nuclear nitrate reductase. *Systematic Botany* **32:** 357–365.

Linder CR, Rieseberg LH. 2004. Reconstructing patterns of reticulate evolution in plants. *American Journal of Botany* **91:** 1700–1708.

Linnaeus C. 1753. Species Plantarum. Stockholm.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* **24:** 2542–2543.

Liu L, Pearl DK, Brumfield RT, Edwards SV. 2008. Estimating species trees using multiple-allele DNA sequence data. *Evolution* **62:** 2080–2091.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evoutionary Biology* **10:** 302.

Liu L, Yu L, Pearl DK, Edwards SV. 2009a. Estimating species phylogenies using coalescence times among sequences. *Systematics Biology* **58:** 468–477.

Liu L, Yu LL. 2011. Estimating species trees from unrooted gene trees. *Systematics Biology* **60:** 661–667.

Liu L, Yu LL, Kubatko L, Pearl DK, Edwards SV. 2009b. Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution* **53:** 320–328.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* **46:** 523–536.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematics Biology* **55:** 21–30.

Maliouchenko O, Palmé AE, Buonamici A, Vendramin GG, Lascoux M. 2007. Comparative phylogeography and population structure of European *Betula* species, with particular focus on *B. pendula* and *B. pubescens*. *Journal of Biogeography* **34:** 1601–1610.

Mallet J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution* **20:** 229–237.

Mallet J. 2007. Hybrid speciation. *Nature* **446:** 279–283.

Marini MA, Barbet-Massin M, Martinez J, Prestes NP, Jiguet F. 2010. Applying ecological niche modelling to plan conservation actions for the red-spectacled Amazon (*Amazona pretrei*). *Biological Conservation* **143:** 102–112.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution* **1:** vev003 doi: 010.1093/ve/vev1003.

Martin DP, Williamson C, Posada D. 2005. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21:** 260–262.

Martin NH, Bouck AC, Arnold ML. 2006. Detecting adaptive trait introgression between Iris fulva and I. brevicaulis in highly selective field conditions. *Genetics* **172:** 2481–2489.

Mayr E 1942. *Systematics and the origin of species.* New York: Columbia University Press.

Mayr E 1963. *Animal species and evolution.* Cambridge, MA: Belknap Press.

McAllister HA, Rushforth K. 2011. *Betula ashburneri*. *Curtis's Botanical Magazine* **28:** 111–118.

McIntosh EJ, Rossetto M, Weston PH, Wardle GM. 2014. Maintenance of strong morphological differentiation despite ongoing natural hybridization between sympatric species of *Lomatia* (Proteaceae). *Annals of Botany* **113:** 861–872.

Meijaard E. 2003. Mammals of south-east Asian islands and their Late Pleistocene environments. *Journal of Biogeography* **30:** 1245–1257.

Mellersh C, Sampson J. 1993. Simplifying detection of microsatellite length polymorphisms. *Biotechniques* **15:** 582–584.

Miller CR, Joyce P, Waits LP. 2002. Assessing allelic dropout and genotyping reliability using maximum likelihood. *Genetics* **160:** 357–366.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17:** 240–248.

Moore WS. 1977. An evaluation of narrow hybrid zones in vertebrates. *The Quarterly Review of Biology* **52:** 263–277.

Moritz C, Faith DP. 1998. Comparative phylogeography and the identification of genetically divergent areas for conservation. *Molecular Ecology* **7:** 419–429.

Muir G, Schlotterer C. 2005. Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.). *Molecular Ecology* **14:** 549–561.

Nagamitsu T, Kawahara T, Kanazashi A. 2006. Endemic dwarf birch *Betula apoiensis* (Betulaceae) is a hybrid that originated from *Betula ermanii* and *Betula ovalifolia*. *Plant Species Biology* **21:** 19–29.

Navarro E, Bousquet J, Moiroud A, Munive A, Piou D, Normand P. 2003. Molecular phylogeny of *Alnus* (Betulaceae), inferred from nuclear ribosomal DNA ITS sequences. *Plant and Soil* **254:** 207–217.

Nichols R. 2001. Gene trees and species trees are not the same. *Trends in Ecology & Evolution* **16:** 358–364.

Nielsen EE, Nielsen PH, Meldrup D, Hansen MM. 2004. Genetic population structure of turbot (*Scophthalmus maximus* L.) supports the presence of multiple hybrid zones for marine fishes in the transition zone between the Baltic Sea and the North Sea. *Molecular Ecology* **13:** 585–595.

Obermayer R, Leitch IJ, Hanson L, Bennett MD. 2002. Nuclear DNA C-values in 30 species double the familial representation in Pteridophytes. *Annals of Botany* **90:** 209–217.

Olsen KM, Schaal BA. 1996. Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. **96:** 5586–5591.

Olszewska MJ, Osiecka R. 1984. The relationship between 2C DNA content, life cycle type, systematic position, and the level of DNA endoreplication in nuclei of parenchyma cells during growth and differentiation of roots in some monocotyledonous species. *Biochemie und Physiologie der Pflanzen* **177:** 319–336.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* **34:** 401–437.

Pálsson S, Lascoux M, Anamthawat-Jónsson K. 2010. Introgression and phylogeography of *Betula nana* (diploid), *B. pubescens* (tetraploid) and their triploid hybrids in Iceland inferred from cpDNA haplotype variation. *Journal of Biogeography* **37:** 2098–2110.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265:** 218–225.

Palmé AE, Su Q, Pálsson S, Lascoux M. 2004. Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among *Betula pendula, B. pubescens* and *B. nana*. *Molecular Ecology* **13:** 167–178.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5:** 568–583.

Pante E, Abdelkrim J, Viricel A, Gey D, France SC, Boisselier MC, Samadi S. 2015. Use of RAD sequencing fro delimiting species. *Heredity* **114:** 450–459.

Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, McMillan WO, Jiggins CD. 2012. Adaptive introgression across species boundaries in *Heliconius* butterflies. *Plos Genetics* **8**.

Pelham J, Gardiner AS, Smith RI, Last FT. 1988. Variation in *Betula pubescens* Ehrh. (Betulaceae) in Scotland: its nature and association with environmental factors *Botanical Journal of the Linnean Society* **96:** 217–234.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* **164:** 10–15.

Pemberton JM, Slate J, Bancroft DR, Barrett JA. 1995. Nonamplifying alleles at microsatellite loci - a caution for parentage and population studies. *Molecular Ecology* **4:** 249–252.

Peterson AT. 2003. Predicting the geography of species' invasions via ecological niche modeling. *The Quarterly Review of Biology* **78:** 419–433.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *Plos One* **7:** e37135.

Petit C, Bretagnolle F, Felber F. 1999. Evolutionary consequences of diploid-polyploid hybrid zones in wild species. *Trends in Ecology & Evolution* **14:** 306–311.

Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R, Ennos R, *et al*. 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300:** 1563–1565.

Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A. 2004. Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161:** 151–164.

Petit RJ, Pineau E, Demesure B, Bacilieri R, Ducousso A, Kremer A. 1997. Chloroplast DNA footprints of postglacial recolonization by oaks. *Proceedings of the National Academy of Sciences of the United States of America* **94:** 9996–10001.

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190:** 231–259.

Phillips SJ, Dudík M, Schapire RE. 2004. A maximum entropy approach to species distribution modeling. *ACM International Proceedings Series* **69:** 655.

Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, Wiens M, Alié A, Morgenstern B, Manuel M, Wörheide G. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Molecular Biology and Evolution* **27:** 1983–1987.

Pinheiro J, Bates D 2000. *Mixed-Effects Models in S and S-PLUS*: Springer, New York.

Plotner J, Uzzell T, Beerli P, Spolsky C, Ohst T, Litvinchuk SN, G-D. G, Reyer H-U, *et al.* 2008. Widespread unidirectional transfer of mitochondrial DNA: a case in western Palaearctic water frogs. *Journal of Evolutionary Biology* **21:** 668–681.

Polz MF, Alm EJ, Hanage WP. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics* **29:** 170–175.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **98:** 13757–13762.

Prentis PJ, White EM, Radford IJ, Lowe AJ, Clarke AR. 2007. Can hybridization cause local extinction: a case for demographic swamping of the Australian native *Senecio pinnatifolius* by the invasive *Senecio madagascariensis*? *New Phytologist* **176:** 902–912.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE. 2014. Demystifying the RAD fad. *Molecular Ecology* **23:** 5937–5942.

Pyakurel A, Wang JR. 2013. Leaf morphological variation among paper birch (*Betula papyrifera* Marsh.) genotypes across Canada. *Open Journal of Ecology* **3:** 284–295.

Qiu YX, Fu CX, Comes HP. 2011. Plant molecular phylogeography in China and adjacent regions: tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. *Molecular Phylogenetics and Evolution* **59:** 225–244.

R Develoment Core Team. 2012. *R:* a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey J, Schemske DW. 2002. Neopolyploidy in flowering plants. *Annual Review of Ecology and Systematics* **33:** 589–639.

Ran JH, Wei XX, Wang XQ. 2006. Molecular phylogeny and biogeography of Picea (Pinaceae): implications for phylogeographical studies using cytoplasmic haplotypes. *Molecular Phylogenetics and Evolution* **41:** 405–419.

Randi E. 2008. Detecting hybridization between wild species and their domesticated relatives. *Molecular Ecology* **17:** 285–293.

Razafimandimbison SG, Kellogg EA, Bremer B. 2004. Recent origin and phylogenetic utility of divergent ITS putative pseudogenes: a case study from Naucleeae (Rubiaceae). *Systematics Biology* **53:** 177–192.

Regel E. 1865. Bemerkungen über die Gattungen *Betula* und *Alnus* nebst Beschreibung einiger neuer Arten. *Bulletin of Society of Naturalist (Moscou)* **38:** 388–434.

Rhymer JM, Simberloff D. 1996. Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics* **27:** 83–109.

Rich TCG, Jermy AC 1998. *Plant Crib.* London: BSBI.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution* **16:** 351–358.

Rieseberg LH, Whitton J, Gardner K. 1999. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* **152:** 713–727.

Rissler LJ, Apodaca JJ. 2007. Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). *Systematic Biology* **56:** 924–942.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61:** 539–542.

Rosenberg NA. 2004. DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4:** 137–138.

Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *Plos One* **7:** e33394.

Sakaguchi S, Bowman DMJS, Prior LD, Crisp MD, Linde CC, Tsumura Y, Isagi Y. 2013. Climate, not Aboriginal landscape burning, controlled the historical demography and distribution of fire-sensitive conifer populations across Australia. *Proceedings of the Royal Society B-Biological Sciences* **280:** 20132182.

Salminen MO, Carr JK, Burke DS, Mccutchan FE. 1995. Identification of breakpoints in Intergenotypic recombinants of HIV type-1 by Bootscanning. *Aids Research and Human Retroviruses* **11:** 1423–1425.

Salzburger W, Baric S, Sturmbauer C. 2002. Speciation via introgressive hybridization in East African cichlids? *Molecular Ecology* **11:** 619–625.

Sampson JF, Byrne M. 2012. Genetic diversity and multiple origins of polyploid *Atriplex nummularia* Lindl. (Chenopodiaceae). *Biological Journal of the Linnean Society* **105:** 218–230.

Sankararaman S, Patterson N, Li H, Paabo S, Reich D. 2012. The date of Interbreeding between Neandertals and modern humans. *Plos Genetics* **8:** e1002947.

Schenk MF, Thienpont CN, Koopman WJM, Gilissen LJWJ, Smulders MJM. 2008. Phylogenetic relationships in *Betula* (Betulaceae) based on AFLP markers. *Tree Genetics & Genomes* **4:** 911–924.

Schmitt T, Seitz A. 2001. Allozyme variation in *Polyommatus coridon* (Lepidoptera: Lycaenidae): identification of ice-age refugia and reconstruction of post-glacial expansion. *Journal of Biogeography* **28:** 1129–1136.

Schoener TW, Gorman GC. 1968. Some niche differences in three Lesser Antillean lizards of the genus *Anolis*. *Ecology* **49:** 819–830.

Scriber JM. 2011. Impacts of climate warming on hybrid zone movement: Geographically diffuse and biologically porous "species borders" *Insect Science*, **18:** 121–159.

Seehausen O. 2004. Hybridization and adaptive radiation. *Trends in Ecology & Evolution* **19:** 198–207.

Selkoe KA, Toonen RJ. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9:** 615–629.

Shaw K, Stritch L, Rivers M, Roy S, Wilson B, Govaerts R. 2014. The red list of Betulaceae. BGCI. Richmond. UK.

Shaw TI, Ruan Z, Glenn TC, Liu L. 2013. STRAW: species tree analysis web server. *Nucleic Acids Research* **41:** 238–241.

Simpson GG. 1951. The species concept. *Evolution* **5:** 285–298.

Simpson GG 1961. *Principles of animal taxonomy.* New York: Columbia University Press.

Skvortsov AK. 2002. A new system of the genus *Betula* L. – the birch. *Bulletin of Moscow Society of Naturalist* **107:** 73–76.

Slatkin M. 1973. Gene flow and selection in a cline. *Genetics* **75:** 733–756.

Šmarda P, Hejcman M, Březinová A, Horová L, Steigerová H, Zedek F, Bureš P, Hejcmanová P, Schellberg J. 2013. Effect of phosphorus availability on the selection of species with different ploidy levels and genome sizes in a long-term grassland fertilization experiment. *New Phytologist* **200:** 911–921.

Smith JM. 1992. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34:** 126–129.

Soltis D, Morris A, McLachlan J, Manos P, Soltis P. 2006. Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology* **15:** 4261–4293.

Soltis DE, Soltis PS. 1999. Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution* **14:** 348–352.

Song KM, Lu P, Tang KL, Osborn TC. 1995. Rapid genome change in synthetic polyploids of Brassica and its Implications for polyploid evolution. *Proceedings of the National Academy of Sciences of the United States of America* **92:** 7719–7723.

Stace CA 2010. *New flora of the British Isles*: Cambridge University Press, Cambridge, U.K.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688–2690.

Stebbins GL. 1959. The role of hybridization in evolution. *Proceedings of the American Philosophical Society* **103:** 231–251.

Stebbins GL 1971. *Chromosomal evolution in higher plants.* London: Edward Arnold.

Stebbins GL. 1984. Polyploidy and the distribution of the arctic–alpine flora: new evidence and a new approach. *Botanica Helvetica* **94:** 1–13.

Stebbins GL. 1985. Polyploidy, hybridization, and the invasion of new habitats. *Annals of the Missouri Botanical Garden* **72:** 824–832.

Stewart JF, Tauer CG, Nelson CD. 2012. Bidirectional introgression between loblolly pine (*Pinus taeda* L.) and shortleaf pine (*P. echinata* Mill.) has increased since the 1950s. *Tree Genetics & Genomes* **8:** 725–735.

Suda J, Meyerson LA, Leitch IJ, Pysek P. 2015. The hidden side of plant invasions: the role of genome size. *New Phytologist* **205:** 994–1007.

Sulkinoja M. 1990. Hybridization, introgression and taxonomy of the mountain birch in SW Greenland compared with related results from Iceland and Finnish Lapland. *Meddelelser om Grönland Bioscience* **33:** 21–29.

Sunnucks P. 2000. Efficient genetic markers for population biology. *Trends in Ecology & Evolution* **15:** 199–203.

Syvanen M. 2012. Evolutionary implications of horizontal gene transfer. *Annual Review of Genetics* **46:** 341–358.

Tanentzap AJ, Zou J, Coomes DA. 2013. Getting the biggest birch for the bang: restoring and expanding upland birchwoods in the Scottish Highlands by managing red deer. *Ecology and Evolution* **3:** 1890–1901.

Tate JA, Simpson BB. 2003. Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. *Systematic Botany* **28:** 723–737.

Templeton AR. 1989. The meaning of species and speciation. In: Otte, D, Endler, JA eds. *Speciation and its consequences.* Sinauer, Sunderland, M. A.

Thórsson Æ, Pálsson S, Lascoux M, Anamthawat-Jónsson K. 2010. Introgression and phylogeography of *Betula nana* (diploid), *B. pubescens* (tetraploid) and

their triploid hybrids in Iceland inferred from cpDNA haplotype variation. *Journal of Biogeography* **37:** 2098–2110.

Thórsson Æ, Salmela E, Anamthawat-Jónsson K. 2001. Morphological, cytogenetic, and molecular evidence for introgressive hybridization in birch. *Journal of Heredity* **92:** 404–408.

Thórsson TH, Pálsson S, Sigurgeirsson A, Anamthawat-Jónsson K. 2007. Morphological variation among *Betula nana* (diploid), *B. pubescens* (tetraploid) and their triploid hybrids in Iceland. *Annals of Botany* **99:** 1183–1193.

Thompson SL, Lamothe M, Meirmans PG, Perinet P, Isabel N. 2010. Repeated unidirectional introgression towards *Populus balsamifera* in contact zones of exotic and native poplars. *Molecular Ecology* **19:** 132–145.

Thomson AM, Dick CW, Dayanandan S. 2015. A similar phylogeographical structure among sympatric North American birches (*Betula*) is better explained by introgression than by shared biogeographical history. *Journal of Biogeography* **42:** 339–350.

Thorn JS, Nijman V, Smith D, Nekaris KAI. 2009. Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: Nycticebus). *Diversity and Distributions* **15:** 289–298.

Tingley MW, Beissinger SR. 2009. Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in Ecology & Evolution* **24:** 625–633.

Toonen RJ, Puritz JB, Forsman ZH, Whitney JL, Fernandez-Silva I, Andrews KR, Bird CE. 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. *Peer Journal* **1:** e203.

Tremblay NO, Schoen DJ. 1999. Molecular phylogeography of Dryas integrifolia: glacial refugia and postglacial recolonization. *Molecular Ecology* **8:** 1187–1198.

Tribsch A, Schonswetter P. 2003. Patterns of endemism and comparative phylogeography confirm palaeoenvironmental evidence for Pleistocene refugia in the Eastern Alps. *Taxon* **52:** 477–497.

Trigo TC, Schneider A, de Oliveira TG, Lehugeur LM, Silveira L, Freitas TRO, Eizirik E. 2013. Molecular data reveal complex hybridization and a cryptic species of Neotropical wild cat. *Current Biology* **23:** 2528–2533.

Trucco F, Tatum T, Rayburn AL, Tranel PJ. 2009. Out of the swamp: unidirectional hybridization with weedy species may explain the prevalence of *Amaranthus tuberculatus* as a weed. *New Phytologist* **184:** 819–827.

Truong C, Palmé AE, Felber F. 2007. Recent invasion of the mountain birch *Betula pubescens* ssp. *tortuosa* above the treeline due to climate change: genetic and ecological study in northern Sweden. *Journal of Evolutionary Biology* **20:** 369–380.

Truong C, Palmé AE, Felber F, Naciri-Graven Y. 2005. Isolation and characterization of microsatellite markers in the tetraploid birch, *Betula pubescens* ssp. *tortuosa*. *Molecular Ecology Notes* **5:** 96–98.

Tsuda Y, Ide Y. 2005. Wide-range analysis of genetic structure of *Betula maximowicziana*, a long-lived pioneer tree species and noble hardwood in the cool temperate zone of Japan. *Molecular Ecology* **14:** 3929–3941.

Tsuda Y, Nakao K, Ide Y, Tsumura Y. 2015. The population demography of *Betula maximowicziana*, a cool-temperate tree species in Japan, in relation to the last glacial period: its admixture-like genetic structure is the result of simple population splitting not admixing. *Molecular Ecology* **24:** 1403–1418.

Twyford AD, Ennos RA. 2012. Next-generation hybridization and introgression. *Heredity* **108:** 179–189.

Vaarama A, Valanne T. 1973. On the taxonomy, biology and origin of *Betula tortuosa* Ledeb. *Reports from the Kevo Subarctic Research Station* **10:** 70–84.

van Dijk T, Noordijk Y, Dubos T, Bink MCAM, Meulenbroek BJ, Visser RGF, van de Weg E. 2012. Microsatellite allele dose and configuration establishment (MADCE): an integrated approach for genetic studies in allopolyploids. *Bmc Plant Biology* **12:25**.

Vekemans X. 2010. What's good for you may be good for me: evidence for adaptive introgression of multiple traits in wild sunflower. *New Phytologist* **187:** 6–9.

Vitousek PM, DAntonio CM, Loope LL, Rejmanek M, Westbrooks R. 1997. Introduced species: a significant component of human-caused global change. *New Zealand Journal of Ecology* **21:** 1–16.

Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, Gignoux C, *et al*. 2013. Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics* **194:** 199–209.

Walters SM. 1968. *Betula* L. in Britain. *Proceedings of the Botanical Society of the British Isles* **7:** 179–180.

Wang B, Climent J, Wang XR. 2015. Horizontal gene transfer from a flowering plant to the insular pine *Pinus canariensis* (Chr. Sm. Ex DC in Buch). *Heredity* **114:** 413–418.

Wang N, Borrell JS, Bodles WJA, Kuttapitiya A, Nichols RA, Buggs RJA. 2014a. Molecular footprints of the Holocene retreat of dwarf birch in Britain. *Molecular Ecology* **23:** 2771–2782.

Wang N, Borrell JS, Buggs RJA. 2014b. Is the Atkinson discriminant function a reliable method for distinguishing between *Betula pendula* and *B. pubescens*? *New Journal of Botany* **4:** 90–94.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW, Pellicer J, Buggs RJA. 2013. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology* **22:** 3098–3111.

Wang S, Meyer E, McKay JK, Matz MV. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods* **9:** 808–810.

Warren DL, Glor RE, Turelli M. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution *Evolution,* **62:** 2868–2883.

Warren DL, Glor RE, Turelli M. 2010. ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography* **33:** 607–611.

Weider LJ, Hobaek A. 2000. Phylogeography and arctic biodiversity: a review. *Annales Zoologici Fennici* **37:** 217–231.

Whitcher IN, Wen J. 2001. Phylogeny and biogeography of *Corylus* (Betulaceae): inferences from ITS sequences. *Systematic Botany* **26:** 283–298.

White TJ, Bruns T, Lee S, Taylor T. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: M. Innis, DG, J. Sninsky, T. White ed. PCR Protocols: A Guide to Methods and Applications. New York: Academic Press, 315–322.

Whitney KD, Randell RA, Rieseberg LH. 2010. Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. *New Phytologist* **187:** 230–239.

Wickham H. 2009. *ggplot2:* elegant graphics for data analysis. 3rd printing 2010 edn, New York: Springer.

Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *Fems Microbiology Reviews* **35:** 957–976.

Wiens JJ. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? *Systematics Biology* **47:** 625–640.

Wiens JJ. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematics Biology* **52:** 528–-538.

Wiens JJ. 2004. The role of morphological data in phylogeny reconstruction. *Systematic Biology* **53:** 653–661.

Wiley EO. 1978. The evolutionary species concept reconsidered. *Systematics Zoology* **27:** 17–26.

Wilsey BJ, Haukioja E, Koricheva J, Sulkinoja M. 1998. Leaf fluctuating asymmetry increases with hybridization and elevation in tree-line birches. *Ecology* **79:** 2092–2099.

Wilsey BJ, Saloniemi I. 1999. Leaf fluctuating asymmetry in tree-line mountain birches, *Betula pubescens* ssp *tortuosa*: genetic or environmentally influenced? *Oikos* **87:** 341–345.

Winkler H. 1904. Betulaceae. *Das Pflanzenreich* **19:** 1–149.

Wolf DE, Takebayashi N, Rieseberg LH. 2001. Predicting the risk of extinction through hybridization. *Conservation Biology* **15:** 1039 – 1053.

Wollenweber E. 1975. Flavonidmuster in Knospenexkret der Betulaceen. *Biochemical Systematics and Ecology* **3:** 47 –52.

Won H, Renner SS. 2003. Horizontal gene transfer from flowering plants to *Gnetum*. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 10824–10829.

Wu CI. 2001. The genic view of the process of speciation. *Journal of Evolutionary Biology* **14:** 851–865.

Yang MA, Malaspinas AS, Durand EY, Slatkin M. 2012. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Molecular Biology and Evolution* **29:** 2987–2995.

Yoo K, Wen J. 2002. Phylogeny and biogeography of *Carpinus* and subfamily Coryloideae (Betulaceae). *International Journal of Plant Sciences* **163:** 641–650.

Zeng J, Li JH, Chen ZD. 2008. A new species of *Betula* section *Betulaster* (Betulaceae) from China. *Botanical Journal of the Linnean Society* **156:** 523–528.

Zeng J, Ren BQ, Zhu JY, Chen ZD. 2014. *Betula hainanensis* (*Betulaster*, Betulaceae), a new species from Hainan Island, China. *Annales Botanici Fennici* **51:** 399–402.

Zhu K, Woodall CW, Clark JS. 2012. Failure to migrate: lack of tree range expansion in response to climate change. *Global Change Biology* **18:** 1042–1052.

Appendix

Publications:

Wang N, McAllister HA, Bartlett PR, Buggs RJA. 2016. Molecular phylogeny and genome size evolution of the genus *Betula* (Betulaceae). *Annals of Botany* Doi: 10.1093/aob/mcw048.

Wang N, Borrell JS, Bodles WJA, Kuttapitiya A, Nichols RA, Buggs RJA. 2014a. Molecular footprints of the Holocene retreat of dwarf birch in Britain. *Molecular Ecology* **23:** 2771–2782.

Wang N, Borrell JS, Buggs RJA. 2014b. Is the Atkinson discriminant function a reliable method for distinguishing between *Betula pendula* and *B. pubescens*? *New Journal of Botany* **4:** 90–94.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW, Pellicer J, Buggs RJA. 2013. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology* **22:** 3098–3111.